
S E
M S M
E S

Proceedings of the

International Congress of Mathematicians

Madrid, August 22–30, 2006

VOLUME II

Invited Lectures

Marta Sanz-Solé
Javier Soria
Juan Luis Varona
Joan Verdera

Editors

European Mathematical Society

Editors:

Marta Sanz-Solé
Facultat de Matemàtiques
Universitat de Barcelona
Gran Via 585
08007 Barcelona
Spain

Javier Soria
Departament de Matemàtica Aplicada i Anàlisi
Facultat de Matemàtiques
Universitat de Barcelona
Gran Via 585
08007 Barcelona
Spain

Juan Luis Varona
Departamento de Matemáticas y Computación
Universidad de La Rioja
Edificio J. L. Vives
Calle Luis de Ulloa s/n
26004 Logroño
Spain

Joan Verdera
Departament de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona)
Spain

2000 Mathematics Subject Classification: 00Bxx

ISBN 978-3-03719-022-7

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available in the Internet at http://dnb.ddb.de.

Contact address:

European Mathematical Society Publishing House
Seminar for Applied Mathematics
ETH-Zentrum FLI C4
CH-8092 Zürich
Switzerland

Phone: +41 (0)44 632 34 36
Email: info@ems-ph.org
Homepage: www.ems-ph.org

Typeset using the author's TEX files: I. Zimmermann, Freiburg
Printed in Germany

9 8 7 6 5 4 3 2 1

Contents

1 Logic and foundations

2 Algebra

3 Number theory

*In case of several authors, invited speakers are marked with an asterisk.

4 Algebraic and complex geometry

5 Geometry

6 Topology

7 Lie groups and Lie algebras

8 Analysis

9 Operator algebras and functional analysis

10 Ordinary differential equations and dynamical systems

Algorithmic randomness and computability

Rod Downey

Abstract. We examine some recent work which has made significant progress in out understanding of algorithmic randomness, relative algorithmic randomness and their relationship with algorithmic computability and relative algorithmic computability.

Mathematics Subject Classification (2000). Primary 68Q30, 68Q15, 03D15, 03D25, 03D28, 03D30.

Keywords. Kolmogorov complexity, computability, degrees of unsolvability, prefix-free complexity, lowness, incompressibility, martingales, computably enumerable.

1. Introduction

In the last few years we have seen some very exciting progress in our understanding of algorithmic randomness and its relationship with computability and complexity. These results have centered around a programme which attempts to answer questions of the following form: when is one real more random than another? How should this be measured? How would such measures of calibration relate to other measures of complexity of reals, such as the traditional measures of relative complexity like Turing degrees, which measure relative computability? These investigations have revealed deep and hitherto unexpected properties of randomness, anti-randomness and algorithmic complexity, as well as pointing at analogs in other areas, and answering questions from apparently completely unrelated areas.

In this paper I will attempt to give a brief (and biased) overview of some of the more recent highlights. I apologize for ignoring important work relating the collection of random strings with complexity theory such as [1], [2], and work on randomness for computably enumerable sets such as Kummer [48], [49], and Muchnik and Positelsky [71], purely for space reasons. This overview will be too short to give a complete account of the all of the progress. For a fuller picture, I refer the reader to the long surveys of Downey, Hirschfeldt, Nies and Terwijn [28], Downey [16], [15], [17], Terwijn [96] and the upcoming monographs Downey and Hirschfeldt [22][1], and Nies [77].

We will look at various methods of calibration by initial segment complexity such as those introduced by Solovay [89], Downey, Hirschfeldt, and Nies [26], Downey,

[1]Available in preliminary form at www.mcs.vuw.ac.nz/~downey.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Hirschfeldt, and LaForte [23], Downey [16], as well as other methods such as lowness notions of Kučera and Terwijn [47], Terwijn and Zambella [97], Nies [75], [76], Downey, Griffiths and Reid [21], and methods such as higher level randomness notions going back to the work of Kurtz [50], Kautz [38], and Solovay [89], and other calibrations of randomness based on changing definitions along the lines of Schnorr, computable, s-randomness, etc. Particularly fascinating is the recent work on lowness, which began with Downey, Hirschfeldt, Nies and Stephan, and developed in a series of deep papers by Nies [75], [76] and his co-authors.

2. Preliminaries

Since most of our results are concerned with effectiveness/computability, we assume that the reader is familiar with the basic facts concerning computability theory/recursion theory. Thus, we will regard countable sets as effectively coded in the natural numbers and consider effective processes on them as computable ones. For example, an effective prediction function would be classified according to it computability. We assume that the reader is also familiar with semi-computable (computably enumerable) processes such as the computably enumerable set coding the halting problem $\emptyset' = \{(e, x) : \text{the } e\text{-th program halts on input } x\}$. Such computable enumerable problems can be represented by sets W defined as $x \in W$ iff $\exists y R(x, y)$, where R is a computable relation. We will call a set in the form $\exists y R(x, y)$, Σ^0_1. If $\mathbb{N} - A$ is Σ^0_1, then we say that A is Π^0_1. If A is both Σ^0_1 and Π^0_1 we say that A is Δ^0_1 (and this is the same as being computable). This process can be extended to the *arithmetical hierarchy*. We will say that A is Σ^0_n iff there is a Π^0_{n-1} relation R such that $x \in A$ iff $\exists y R(x, y)$. (Equivalently, x is in A iff $\exists y \forall z \ldots$ (with n alternations)$Q(x, y, z, \ldots)$ and Q computable.) Analogously, we can define Π^0_n and Δ^0_n. We will also assume that the reader is familiar with the process of *relativization* which means that we put oracles (allowing for "read only memory") on our machines. These oracles allow for computations in which a finite number of effectively generated membership queries of the oracle set are allowed. Thus, for instance, $A' = \{(e, x) : \text{the } e\text{-th program halts on input } x \text{ when given oracle } A\}$. This is the halting problem *relativized to* A, usually pronounced "A-jump". If we classify sets under the preordering $\leq_T$ we will write $A \leq_T B$ to mean that membership of A can be computed by a program with access to B as an oracle. (Here we identify sets with their characteristic functions, and hence as reals: members of Cantor space 2^ω.) The equivalence classes of $\leq_T$, which calibrate countable sets into classes of "equi-computability" are called Turing degrees, after the famous Alan Turing. We remark that the simplest kind of Turing reduction is called an m-reduction (for many-one) and is defined as follows: $A \leq_m B$ means that there is a computable function f such that $x \in A$ iff $f(x) \in B$. Thus to figure out if x is in A from B, the algorithm simply says : compute $f(x)$ and ask B if $f(x)$ is in B. It is easy to show that for any computably enumerable set A, $A \leq_m \emptyset'$, so that the halting problem $\emptyset'$ if m-complete, in that it is the most complicated computably

enumerable set as measured by m-reducibility[2]. We remark that the relativization of the halting problem be algorithmically unsolvable is that $A' \not\leq_T A$ for any set A. The relativization of the halting problem is intrinsically tied with the halting problem. Namely, $\emptyset''$, which is defined as the halting problem gained with the halting problem as an oracle is a natural Σ^0_2 set and it can compute any Π^0_2 set and any Σ^0_2 set, and similarly for $\emptyset^{(n+1)}$.

Any other notions from computability needed are introduced in context. We also refer the reader to Soare [86] for further background material in computability, and to Li–Vitanyi [56] or Calude [6] for general background in algorithmic randomness.

In this paper "real" will be interpreted as a member of Cantor space 2^ω with sub-basic clopen sets $[\sigma] = \{\sigma\alpha : \alpha \in 2^\omega\}$, for $\sigma \in 2^{<\omega}$. This space is equipped with the standard Lebesgue measure, where, for $\sigma \in 2^{<\omega}$, $\mu([\sigma]) = 2^{-|\sigma|}$. There have been investigations on other measures than the uniform one, and on other spaces (the latter notably by Gács [34]), but space precludes a thorough discussion here. For Cantor space up to degree things, speaking loosely, it does not matter measure is used, so long as it is not atomic. Finally, the initial segment of a real α (or a string) of length n will be denoted by $\alpha \upharpoonright n$.

3. Three approaches to randomness

In terms of measure a any two reals occur with probability zero, yet we would argue that a real $\alpha = 01010101\ldots$ would not seem random. How should we understand this?

3.1. Martin-Löf randomness. The first author to attempt to grapple with trying to "define" randomness was von Mises [101]. Von Mises was a statistician and attempted to define randomness in terms of statistical laws. For instance, he argued that a random real should pass all statistical tests. Thus, he argued, if one "selected" from a real $\alpha = a_0a_1\ldots$ some subsequence $a_{i_0}, a_{i_1}, \ldots$, then $\lim_{n\to\infty} \frac{|\{j : a_{i_j} = 1 \wedge 1 \leq j \leq n\}|}{n}$ should be $\frac{1}{2}$. Naturally, von Mises lacked the language needed to suggest which selection rules should be considered. That awaited the development of computable function theory in the 1930s by Church and others, which then allowed us to argue that a random real should be "computably stochastic" in the sense of von Mises.

Unfortunately, Wald and others showed that there are some significant problems (see van Lambalgen [99] for a discussion) with this approach, known as computable stochasticity. Here I refer the reader to Ambos-Spies [3], Merkle [62], [63], and Uspensky, Semenov and Shen [98]. The first really acceptable version of von Mises idea was developed by Per Martin-Löf in [60]. He argued that any effective statistical

[2]Additionally, it might seem that there might be various versions of the halting problem depending of which programming language, or which encoding, is used. It can be shown that that are all of the same m-degree, and hence are basically all the same. More on this in the context of randomness later.

test was an effective null set and a random real should be one that simply avoids any effective null set.

The notion of an effective collection of reals is are called effective classes. As a direct analog of the arithmetical hierarchy. A Σ^0_1 class U is a "c.e. set of reals" in the sense that there is a computable relation R such that for each real α, $\alpha \in U$ iff $\exists x\, R^\alpha(x)$, where R^α denotes R with oracle α. An equivalent definition is that U is a Σ^0_1 class iff there is a c.e. set of intervals W such that $U = \cup\{[\sigma] : \sigma \in W\}$. Now we can make our intuition of avoiding all effective statistical tests more precise, as follows.

Definition 3.1 (Martin-Löf [60]). A set of reals $A \subseteq 2^\omega$ is Martin-Löf null (or Σ_1*-null*) if there is a uniformly c.e. sequence $\{U_i\}_{i\in\omega}$ of Σ^0_1-classes (called a *Martin-Löf test*) such that $\mu(U_i) \leq 2^{-i}$ and $A \subseteq \bigcap_i U_i$. $\alpha \in 2^\omega$ is Martin-Löf random, or 1-random, if $\{\alpha\}$ is not Σ_1-null.

This definition and variations form common bases for the theory of algorithmic randomness. There are also two other approaches aside from the measure-theoretical. These include the incompressibility paradigm and the unpredictability paradigm.

It is possible to calibrate randomness in a method similar to the arithmetical hierarchy, by defining n-randomness exactly as above, except that Σ^0_1 null sets are replaced by Σ^0_n null sets. It can be shown (Kurtz [50]) that $n+1$-randomness is 1-randomness relative to $\emptyset^{(n)}$, Stuart Kurtz [50] was the first meaning that if $\emptyset'$ is given as an oracle, what is the analog of Martin-Löf randomness. to systematically examine the relationship between n-randomness and the computability, although some unpublished work was to be found in Solovay [89], and 2-randomness was already to be found in Gaifman and Snir [35], in implicit form.

There has been quite some work clarifying the relationship between Turing reducibility and n-randomness. For example, it has long been known that if $\mathbf{a}$ is $n+1$-random then $\mathbf{a}$ is GL_n, meaning that $\mathbf{a} \cup \mathbf{0}^n = (\mathbf{a} \cup \mathbf{0})^n$, and that the "almost all" theory of degrees is decidable (Stillwell [93]). Recently some lovely new work has emerged. As an illustration, we mention the following unexpected result.

Theorem 3.2 (Miller and Yu [69]). *Suppose that $A \leq_T B$ and B is n-random and A is* 1*-random. Then A is n-random.*

3.2. Kolmogorov complexity. The key idea here is that a random string (as generated by a coin toss, say) should not be easily described by a short program. Thus, 10^{100} is easily described by a description much shorter than its length. This incompressibility idea was the famous approach pioneered by Kolmogorov [41] (also cf. Solomonoff [88]). For our programming language (which we take as Turing machines) we consider the lengths of strings σ producing a string τ. Think of σ as a description of τ under the action of the machine N. Then the N-complexity of the τ is the *length* of the shortest σ from which N produces τ. Since we can enumerate the machines $M_0, M_1, \ldots$, we can make a universal machine M which acts as

$M(1^{e+1}0\sigma) = M_e(\sigma)$.Thus, there is canonical choice for the choice of machine up to a constant, and we define the (plain) *Kolmogorov complexity* of τ as

$$C(\tau) = \min\{\infty, |\sigma| : M(\sigma) = \tau\}.$$

The we would say that τ is C-random iff $C(\tau) \geq |\tau|$. We will also need conditional versions of this (and other) measures. We will write $C(\sigma|\nu)$ as the conditional plain complexity of σ given ν as an oracle. (We will use analogous notation for K below.)

Plain Kolmogorov complexity produces a nice theory of randomness for strings, but as Martin-Löf argued, plain complexity fails to capture the intentional meaning of "the bits of σ producing the bits of τ". This is the length of σ itself can be used in the program, giving $\tau + |\tau|$ many bits of information. Thus, it is easily shown that if α is sufficiently long then there is some n such that $C(\alpha \upharpoonright n) < n$, meaning that there are *no* random reals if we take randomness to mean that all initial segments should be random[3].

This problem was overcome by Levin [51], [54], Schnorr [84], and Chaitin [10], using monotone, process and prefix-free complexities. Here we focus on the prefix-free complexity. Recall that A of intervals is called *prefix-free* iff for all σ, τ, if $\sigma \prec \tau$, then $[\sigma] \in A$ implies $[\tau] \notin A$. Note that for such a set A,

$$\mu(A) = \sum\{2^{-|\sigma|} : [\sigma] \in A\}.$$

Levin and then Chaitin defined prefix-free Kolmogorov complexity using machines whose domains were prefix free. Again there is a universal one U (same argument) and we define

$$K(\tau) = \min\{|\sigma| : U(\sigma) = \tau\}.$$

Finally we can define a real to be K-random iff for all n, $K(\alpha \upharpoonright n) \geq n - O(1)$. The concepts of Martin-Löf randomness and K-randomness are tied together as follows.

Theorem 3.3 (Schnorr, see Chaitin [10], [12]). *$A \in 2^\omega$ is Martin-Löf random if and only if it is K-random.*

Given Schnorr's Theorem, Solovay had asked if $\liminf_s K(\Omega \upharpoonright n) - n \to \infty$. This was solved affirmatively by Chaitin. However, there is a very attractive generalization of this due to Miller and Yu who show that the complexity of a random real must be above n eventually by "quite a bit."

Theorem 3.4 (Ample Excess Lemma, Miller and Yu [69]). *A real α is random iff*

$$\sum_{n\in\mathbb{N}} 2^{n-K(\alpha|n)} < \infty.$$

[3] Specifically, every string ν corresponds to some number (string) in the length/lexicographic ordering of $2^{<\omega}$. Given a long string α, take any initial segment $\alpha \upharpoonright n$. This corresponds to a number m in this way. Now consider the programme which, on input ρ interprets ρ's length as a string γ and outputs $\gamma\rho$. If this programme is enacted on $\alpha \upharpoonright_{n+1}^{n+m}$ the segment of α of length m beginning after α, it will output $\alpha \upharpoonright n + m$, allowing for compression of arbitrary segments.

Corollary 3.5 (Miller and Yu [70]). *Suppose that f is an arbitrary function with $\sum_{m\in\mathbb{N}} 2^{-f(m)} = \infty$. Suppose that α is 1-random. Then there are infinitely many m with $K(\alpha \upharpoonright m) > m + f(m)$.*

The reader might wonder whether plain complexity could be used to characterize 1-randomness. There had been some natural "C-conditions" which had been shown to guarantee randomness. Martin-Löf showed that if a real had infinitely often maximal C -complexity then it would be random. That is, Kolmogorov observed that the greatest plain complexity a string σ can have is $|\sigma|$. We will say that a real is *Kolmogorov random* iff $\exists^\infty n[C(\alpha \upharpoonright n) = n - O(1)$. If A is Kolmogorov random it is 1-random. But recently more has been shown. Chaitin showed that the highest prefix-free complexity a string can have is $|\sigma| + K(|\sigma|)$, and we define α to be strongly Chaitin random iff $\exists^\infty n[(K(\alpha \upharpoonright n) > n + K(n) - O(1)]$. Solovay [89] (see Yu, Ding, Downey [107]) showed that each 3-random is strongly Chaitin random, and every strongly Chaitin random real is Kolmogorov random and hence 1-random. It is not known if every Kolmogorov random real is strongly Chaitin random. The following remarkable result shows that Kolmogorov randomness can be characterized in terms of the randomness hierarchy.

Theorem 3.6 (Nies, Stephan and Terwijn [78]). *Suppose that α is Kolmogorov random. Then α is 2-random.*

Theorem 3.7 (Miller [66], Nies, Stephan and Terwijn [78]). *A real α is 2-random iff α is Kolmogorov random.*

We remark that there seems no *prima facie* reason for 2-randomness to be the same as Kolmogorov randomness! The question of whether there was a natural condition in terms of plain complexity which characterized 1-randomness was finally solved by Miller and Yu, having been open for 40 years.

Definition 3.8 (Miller and Yu [69]). Define a computable function $G : \omega \to \omega$ by

$$G(n) = \begin{cases} K_{s+1}(t), & \text{if } n = 2^{\langle s,t\rangle} \text{ and } K_{s+1}(t) \neq K_s(t), \\ n, & \text{otherwise.} \end{cases}$$

Theorem 3.9 (Miller and Yu [69]). *For $x \in 2^\omega$, the following are equivalent:*

(i) *x is 1-random.*

(ii) (One direction of this is in Gács [32]) $(\forall n)\ C(x \upharpoonright n) \geq n - K(n) \pm O(1)$.

(iii) $(\forall n)\ C(x \upharpoonright n) \geq n - g(n) \pm O(1)$, *for every computable $g : \omega \to \omega$ such that $\sum_{n\in\omega} 2^{-g(n)}$ is finite.*

(iv) $(\forall n)\ C(x \upharpoonright n) \geq n - G(n) \pm O(1)$.

While it is not hard to show that almost all reals are random (as one would hope), Schnorr's Theorem allows us to easily show that there are explicit random reals. The halting probabilities of prefix-free Turing machines occupy the same place in algorithmic randomness as computably enumerable sets (the domains of partial computable functions) do in classical computability theory. They are called *left-computably enumerable reals* (left-c.e.) and are defined as the limits of increasing computable sequences of rationals. A special left-c.e. real is $\Omega = \sum_{U(\sigma)\downarrow} 2^{-|\sigma|}$ where U is a universal prefix free machine.

Theorem 3.10 (Chaitin [10], [12]). *Ω is Martin-Löf random.*

Chaitin's Ω has had a lot of popular attention. It allows us to prove Gödel's incompleteness theorem and the like using Kolmogorov complexity. Solovay [89] was the first to look at basic computability-theoretical aspects of Ω. For instance, consider $D_n = \{x : |x| \leq n \wedge U(x)\downarrow\}$. Solovay proved that $K(D_n) = n + O(1)$, where $K(D_n)$ is the K-complexity for an index for D_n. Solovay also proved the following basic relationships between D_n and $\Omega \upharpoonright n$.

Theorem 3.11 (Solovay [89]).

(i) $K(D_n|\Omega \upharpoonright n) = O(1)$[4].

(ii) $K(\Omega \upharpoonright n|D_{n+K(n)}) = O(1)$.

The reader should note that in classical computability theory, we usually talk of *the* halting problem, whereas here the definition of Ω seems thoroughly machine dependent. To try to address this issue, Solovay [89] introduced the following definition, which is a kind of analytic version of m-reducibility.

Definition 3.12 (Solovay [89]). We say that a real α is *Solovay reducible* to β (or β *dominates* α), $\alpha \leq_S \beta$, iff there is a constant c and a partial computable function f, so that for all $q \in \mathbb{Q}$, with $q < \beta$,

$$c(\beta - q) > \alpha - f(q).$$

The intuition here is a sequence converging to β can generate one converging to α at the same rate, as clarified by Calude, Hertling, Khoussainov, Wang [9]. It is easy to see that $\leq_S$ implies $\leq_T$ for reals. Since there are only $O(2^{2d})$ many reals within a radius of 2^{-n+d} of a string representing a rational whose dyadic expansion has length n, it follows that $\leq_S$ has the *Solovay Property* of the lemma below.

Lemma 3.13 (Solovay [89]). *If $\alpha \leq_S \beta$ then there is a c such that, for all n,*

$$K(\alpha \upharpoonright n) \leq K(\beta \upharpoonright n) + c.$$

The same also holds for C in place of K.

[4]Indeed, $D_n \leq_{wtt} \Omega \upharpoonright n$ via a weak truth table reduction with identity use, where a Turing reduction is a weak truth table one if there is a computable bound on the size of the queries used.

This lemma shows that, if $\Omega \leq_S \beta$, then β is Martin-Löf random. The next result says the being Ω-like means that a left-c.e. real look like Ω.

Theorem 3.14 (Calude, Hertling, Khoussainov, Wang [9]). *Suppose that β is a left-c.e. real and that $\Omega \leq_S \beta$. Then β is a halting probability. That is, there is a universal machine $\widehat{U}$ such that $\mu(\mathrm{dom}(\widehat{U})) = \beta$.*

The final piece of the puzzle was provided by the following lovely result of Kučera and Slaman.

Theorem 3.15 (Kučera and Slaman [46]). *Suppose that α is random and a left-c.e. real. Then for all left-c.e. reals β, $\beta \leq_S \alpha$, and hence α is a halting probability.*

We know that all reals have complexity oscillations. The Kučera–Slaman Theorem says that for left-c.e. random reals, they all happen in the same places. Downey, Hirschfeldt and Nies [26], and Downey, Hirschfeldt and LaForte [24] were motivated to look at the structure of computably enumerable reals under Solovay reducibility. The structure remains largely unexplored.

Theorem 3.16 (Downey, Hirschfeldt and Nies [26]).

(i) *The Solovay degrees of left-c.e. reals forms a distributive upper semilattice, where the operation of join is induced by $+$, arithmetic addition* (*or multiplication*) (*namely* $[x] \vee [y] \equiv_S [x + y]$).

(ii) *This structure is dense.*[5] *In fact if* $\mathbf{a} < \mathbf{b} < [\Omega]$ *then there exist incomparable* $\mathbf{b_1}, \mathbf{b_2}$ *with* $\mathbf{a} < \mathbf{b_1} \vee \mathbf{b_2} = \mathbf{b}$.

(iii) *However, if* $[\Omega] = \mathbf{a} \vee \mathbf{b}$ *then either* $[\Omega] = \mathbf{a}$ *or* $[\Omega] = \mathbf{b}$.

Theorem 3.17 (Downey and Hirschfeldt [22]). *There exist left-c.e.* sets *A and B such that the Solovay degrees of A and B have no infimum in the* (*global*) *Solovay degrees.*

Theorem 3.18 (Downey, Hirschfeldt, and LaForte [24]). *The first order theory of the uppersemilattice of the Solovay degrees of left-c.e. reals is undecidable.*

We can view Ω as a fundamental operator on reals in the same way as we do for the jump operator. However, we need real care when dealing with relativizing Ω. We will take the notion of *universal machine* to mean that the machine U should be universal (and hence prefix-free) for all oracles, and if M_e is any machine, then M_e should be effectively coded in U, meaning that for some τ, $M_e(\sigma) = U(\tau\sigma)$. This definition avoids pathological machines.

The properties of omega operators acting on Cantor space and their relationship with, for instance, Turing reducibility was really initiated by Downey, Hirschfeldt, Miller and Nies [25]. It had been hoped, for instance, that these might be degree invariant operators on 2^ω. This hope failed about as badly as it could.

[5] In fact, Downey and Hirschfeldt [22] have sown the Density Theorem holds for the left-c.e. reals for any measure of relative randomness which has a Σ^0_3 definition, has a top degree of $[\Omega]$, + is a join, and where the computable sets are in the zero degree.

Theorem 3.19 (Downey, Hirschfeldt, Miller, Nies [25]). *For any omega operator* Ω*, there are reals* $A =^* B$ (*meaning that they differ only by a finite amount*) *such that* Ω^A *and* Ω^B *are relatively random* (*and hence* $\Omega^A|_T \Omega^B$).

One the other hand, omega operators do have some fascinating properties.

Theorem 3.20 (Downey, Hirschfeldt, Miller, Nies [25]). *Omega operators are lower semicontinuous but not continuous, and moreover, that they are continuous exactly at the* 1*-generic reals*[6].

In some sense Ω is kind of a red herring amongst random reals. It gives the impression that random reals have high computational power. Also results such as the famous Kučera–Gács Theorem below say that some random reals have high computational power.

Theorem 3.21 (Kučera [42], Gács [33]). *Every set is Turing* (wtt-)*reducible to a Martin-Löf random set.*

We remark that it is by no means clear this result should be true. After all, the very first result connecting measure and computability was the following:

Theorem 3.22 (de Leeuw, Moore, Shannon, and Shapiro [14]). *Define the* enumeration probability *of A as*

$$P(A) = \mu(\{X \in 2^\omega : U^X = A\}),$$

where U is some universal machine. Then if $P(A) > 0$*, A is a computably enumerable set.*

An immediate corollary is the result first stated by Sacks [81] that A is computable iff $\mu(\{Y : A \leq_T Y\}) > 0$.

The question is : "How do we reconcile the notions of high computational power and high randomness?". Frank Stephan gave a clarification to this dichotomy. We say that a function f is fixed point free iff for all partial computable functions φ_e, $f(e) \neq \varphi_e(e)$. We will say a set A has PA if it has the computational power to compute $\{0, 1\}$ valued fixed point free function[7]. Whilst Kučera [44], [45] had shown that random reals can always compute fixed point free functions[8], Stephan showed that the randoms above the degree of the halting problem are the only ones with sufficient computational power to be able to compute a $\{0, 1\}$-valued one[9].

Theorem 3.23 (Stephan [91]). *Suppose that* $\mathbf{a}$ *is PA and* 1*-random. Then* $\mathbf{0}' \leq_T \mathbf{a}$.

[6] Here recall that x is 1-generic means that it is Cohen generic for 1 quantifier arithmetic.

[7] They are called PA degrees since the coincide with the degrees bounding complete extensions of Peano Arithmetic. (Scott [85], Solovay.)

[8] Additionally, Kučera proved that if A is n-random, then A bounds an n-FPF function. We refer the reader to [45] or [22] for definitions and details.

[9] Also, Kjos-Hanssen, Merkle, and Stephan [39] give a variant of in terms of Kolmogorov complexity and is in some sense an explanation why it is true.

All of this might leads the reader to guess that Ω, and hence all halting probabilities, have little to do with algorithmic randomness in general. Again this is not the case.

Theorem 3.24 (Downey, Hirschfeldt, Miller, Nies [25]). *Suppose that A is 2-random. Then there is a universal machine U and set B such that $A = \Omega_U^B$.*

That is, almost all randoms are halting probabilities. Notice that $\overline{\Omega}$ is random, but cannot be a halting probability relative to any oracle.

By analyzing the "majority vote" proof of Sacks Theorem, it is easy to show that if A is 2-random and $B \leq_T A$, then A is *not* random relative to B. Thus Theorem 3.24 stands in contrast the classical theorem from Kurtz' regrettably unpublished Thesis. (Proofs of this result and others from Kurtz's Thesis, and from Solovay's notes can be found in Downey and Hirschfeldt [22].)

Theorem 3.25 (Kurtz [50]). *Suppose that A is 2-random. Then there is a set $B \leq_T A$ such that A is computably enumerable relative to B.*

3.3. Martingales and the prediction paradigm. The last major approach to the concept of algorithmic randomness uses the intuition that random reals should be *hard to predict*. This can be formalized by imagining you had some "*effective*" betting strategy which worked on the bits of a real α. At each stage you get to try to predict the next bit of α, knowing the previous n bits. This idea leads to the following concept.

Definition 3.26 (Levy [55]). A *martingale* (*supermartingale*) is a function $f : 2^{<\omega} \mapsto \mathbb{R}^+ \cup \{0\}$ such that for all σ,

$$f(\sigma) = \frac{f(\sigma 0) + f(\sigma 1)}{2} \quad (\text{resp.} f(\sigma) \geq \frac{f(\sigma 0) + f(\sigma 1)}{2}).$$

We say that the (super-)martingale *succeeds* on a real α, if $\limsup_n F(\alpha \upharpoonright n) \to \infty$.

Martingales were introduced by Levy [55], and Ville [102] proved that null sets correspond to success sets for martingales. They were used extensively by Doob in the study of stochastic processes. Schnorr [82], [83] effectivized the notion of a (super-)martingale.

Definition 3.27. We will define a (super-)martingale f as being *effective* or *computably enumerable* if $f(\sigma)$ is a c.e. real, and at every stage we have effective approximations to f in the sense that $f(\sigma) = \lim_s f_s(\sigma)$, with $f_s(\sigma)$ a computable increasing sequence of rationals.

We remark that the reader might have expected that an effective martingale would be one with f a computable function rather than one with computable *approximations*. This is an important point and we return to it later.

Theorem 3.28 (Schnorr [82]). *A real α is Martin-Löf random iff no effective (super-) martingale succeeds on α.*

Thus, we have nice evidence that we have captured a reasonable notion of algorithmic randomness in that the three approaches, measure-theoretical, compressional, and predictability, all give the same class.

3.4. Schnorr's critique. In [82], [83], Schnorr argued that Theorem 3.28 demonstrated a clear failure of the intuition behind the definition of algorithmic randomness in that it we had *computable enumerable* betting strategies corresponding to Martin-Löf randomness rather than *computable* ones. Schnorr proposed the two variations below, and these have had attracted considerable interest recently. The first is to replace computably enumerable martingales by computable martingales and obtain the concept of *computably random* meaning that no computable martingale can succeed on the real. The second is to take the definition of Martin-Löf randomness (Definition 3.1) and replace $\mu(U_i) \leq 2^{-i}$ by $\mu(U_i) = 2^{-i}$ so that we know exactly the measure of the test sets, and hence can decide if $[\sigma] \in U_i$ by waiting until we know the measure of U_i to within $2^{-|\sigma|}$. Some clarification of the relationships between these two concepts was obtained by Schnorr.

Definition 3.29. We say that a computable martingale *strongly* succeeds on a real x iff there is a computable unbounded nondecreasing function $h\colon \mathbb{N} \mapsto \mathbb{N}$ such that $F(x \upharpoonright n) \geq h(n)$ infinitely often.

Theorem 3.30 (Schnorr [82]). *A real x is Schnorr random iff no computable martingale strongly succeeds on x.*

Thus Martin-Löf randomness implies computable randomness which implies Schnorr randomness. None of the implications can be reversed (van Lambalgen [99]). These concepts were somewhat ignored for maybe 20 years after Schnorr defined them, possibly because Martin-Löf randomness sufficed for many tasks, and because they were rather more difficult to handle. There are no universal tests, for instance, for Schnorr randomness. Recently, Downey and Griffiths [19] gave a nice characterization of Schnorr randomness in terms of *computable* machines. Here prefix-free M is called computable iff the measure of its domain is a computable real.

Theorem 3.31 (Downey and Griffiths [19]). *A real α is Schnorr random iff for all computable machines M, there is a constant c such that, for all n, $K_M(\alpha \upharpoonright n) \geq n - c$.*

Related here is yet another notion of randomness called Kurtz or weak randomness. We define a *Kurtz test* (resp. Kurtz n-test) to be a Σ^0_1 (resp. Σ^0_n-) class of measure 1. Then a real A is called *weakly* (n-)*random* or *Kurtz n-random*[10] if it passes all Kurtz (n-)tests, meaning that $A \in U$ for all such U. There is a null test version.

[10] Now it could be argued that weak randomness is not really a randomness notion at all, but rather a genericity notion. However, for $n \geq 2$ it is certainly a randomness notion, and $n = 2$ corresponds to "Martin-Lóf tests with no effective rate of convergence."

Definition 3.32 (Wang [103]). A *Kurtz null test* is a collection $\{V_n : n \in \mathbb{N}\}$ of c.e. open sets, such that

(i) $\mu(V_n) \leq 2^{-n}$, and

(ii) there is a computable function $f : \mathbb{N} \to (\Sigma^*)^{<\omega}$ such that $f(n)$ is a canonical index for a finite set of σ's, say, $\sigma_1, \ldots, \sigma_n$ and $V_n = \{[\sigma_1], \ldots, [\sigma_n]\}$.

Theorem 3.33 (Wang [103], after Kurtz [50]). *A real α is Kurtz random iff it passes all Kurtz null tests.*

Wang also gave a martingale version of Kurtz randomness.

Theorem 3.34 (Wang [103]). *A real α is Kurtz random iff there is no computable martingale F and nondecreasing computable function h, such that for* almost all *n,*

$$F(\alpha \upharpoonright n) > h(n).$$

This should be directly compared with Schnorr's characterization of Schnorr randomness in terms of martingales and computable orders. Downey, Griffith and Reid [21] gave a machine characterization of Kurtz randomness, and showed that each computably enumerable non-zero degree contained a Kurtz random left-c.e. real. This contrasted with the theorem of Downey, Griffiths and LaForte [20] who showed that if a left-c.e. real was Kurtz random, then its Turing degree must resemble the halting problem in that it must be high (i.e. $A' \equiv_T \emptyset'$). The definitive (and rather difficult) result here is the following which builds on all of this work.

Theorem 3.35 (Nies, Stephan and Terwijn [78]). *For every set A, the following are equivalent.*

(I) *A is high (i.e. $A' \geq_T \emptyset''$).*

(II) *There exists $B \equiv_T A$, such that B is computably random but not Martin-Löf random.*

(III) *There exists $C \equiv_T A$, such that C is Schnorr random but not computably random.*

Moreover, the examples can be chosen as left-c.e. reals if the degrees are computably enumerable.

Remarkably, outside of the high degrees the notions coincide.

Theorem 3.36 (Nies, Stephan and Terwijn [78]). *Suppose that a set A is Schnorr random and does not have high degree. Then A is Martin-Löf random.*

An even more unexpected collapse occurs for the special class of degrees called hyperimmune-free degrees. Following Miller and Martin [73], we say that A is *hyperimmune-free* iff for all functions $f \leq_T A$, there is a computable function g such that for all x, $f(x) \leq g(x)$.

Theorem 3.37 (Nies, Stephan, Terwijn [78]). *Suppose that A is of hyperimmune-free degree. Then A is Kurtz random iff A is Martin-Löf random.*

Space precludes me for discussing a very attractive possible refutation of Schnorr's critique proposed by Muchnik, Semenov, and Uspensky [72] who looked at *nonmonotonic* betting strategies, where now we no longer pick the bits of the real in order. The open question is whether using computable nonmonotonic supermartingales, we might capture the notion of Martin-Löf randomness. We refer the reader to the paper of Merkle, Miller, Nies, Reimann and Stephan [65] and [72].

3.5. Hausdorff dimension. Whilst I do not really have enough space to do justice to the area, there has been a lot of very interesting work concerning effective Hausdorff dimension of even single reals and strings. For instance, we would expect that if $\Omega = w_0 w_1 \ldots$ then somehow $w_0 00 w_1 00 w_2 00 \ldots$ should be "$\frac{1}{3}$ random." We can address this issue using a refinement of the class of measure zero sets is given by the theory of Hausdorff Dimension. In 1919 Hausdorff [36] generalized earlier work of Carathéodory to define a notion of an s-dimensional measure to include non-integer values. The basic idea is that you replace measure by a kind of generalized measure, where $\mu([\sigma])$ is replaced by $2^{-s|\sigma|}$ where $0 < s \leq 1$. With $s = 1$ we get normal Lebesgue measure. For $s < 1$ we get a refinement of measure zero. We can translate this cover version into a s-gale (a version of martingales, namely $f(\sigma) = 2^{-s}(f(\sigma 0) + f(\sigma 1))$) definition in the same way that it is possible to frame Lebesgue measure in terms of martingales.

Here we are viewing betting strategies in a *hostile environment* (a model of Jack Lutz), where "inflation" is acting so *not winning* means that we automatically lose money. (For normal martingales, we are to choose not to bet on some bit saving our funds for later bits and this has no effect. Here failing to bet means that our capital shrinks. The most hostile environment where we can win will be the effective Hausdorff dimension.) That is, roughly speaking, it can be shown that there is some limsup where the s-measure is not zero, and this is called the Hausdorff dimension of the set.

The study of effective dimension was pioneered through the work of Jack Lutz though as with much of the area of randomness there is a lot of history. In any case, for the effective version through the work of Lutz, Mayordomo, Hitchcock, Staiger and others we find that the notion corresponds to $\liminf_n \frac{K(A\upharpoonright n)}{n}$, and can take that as a *working definition* of effective Hausdorff dimension. (Here I must refer the reader to Lutz [58], [59] for more details and history.)

With this definition, it can easily be shown that the "00" version of Ω above really has Hausdorff dimension $\frac{1}{3}$ and in fact is $\frac{1}{3}$ random as in Tadaki [94].

Terwijn [95], [96] and Reimann [80] have very nice results here relating Hausdorff dimension to degree structures. The latter as well and Lutz and Mayordomo have also looked at other dimensions, such as effective packing dimension, which can be characterized as $\limsup_n \frac{K(A\upharpoonright n)}{n}$. Again it is possible to examine these concepts for

stronger and weaker randomness notions such as Schnorr dimension. For instance, Downey, Merkle and Reimann [30] have shown that it is possible to have computably enumerable *sets* with nonzero Schnorr packing dimension, whereas their Schnorr Hausdorff dimension is 0. Much work remains to be done here with a plethora of open questions.

We finish this section by remarking that Lutz [58], [59] has even developed a notion of dimension for individual *strings*. The approach is to replace s-gales by "termgales" which are the analogues of s-gales for terminated strings. In essence he has characterized dimension for individual strings exactly in terms of prefix-free Kolmogorov complexity. Space does not allow for the development of this theory and we refer the reader to Lutz [58], [59] or Downey and Hirschfeldt [22] for further details.

4. Calibrating randomness

We have seen that we can classify randomness in terms of initial segment complexity. Thus it seems reasonable to think that we should also be able to classify *relative* randomness in terms of *relative* initial segment complexity. This motivates the following definition.

Definition 4.1 (Downey, Hirschfeldt, and LaForte [23]). We say a pre-ordering $\preceq$ is am *Q-initial segment measure of relative randomness* iff it obeys the Solovay property met earlier: $A \preceq B$ means that for all n, $Q(A \upharpoonright n) \leq Q(B \upharpoonright n) + O(1)$.

Here we are thinking of Q as C or K. We have already seen that Solovay reducibility is a measure of relative randomness and can be used to characterize the left-c.e. random reals. However, Solovay reducibility has a number of limitations such as being too fine and only really relating to left-c.e. reals.

There are a number of other interesting measures of relative randomness. They include segment ones $\leq_C$ and $\leq_K$ which are defined in the obvious way. Others include the following introduced by Downey, Hirschfeldt and LaForte [23]:

(i) $A \leq_{sw} B$ iff there is a c and a wtt procedure Γ with use $\gamma(n) = n + c$, and $\Gamma^B = A$. If $c = 0$, then this is called *ibT-reducibility* and is the one used by Soare and Csima in differential geometry, such as Soare [87].

(ii) $A \leq_{rK} B$ means that there is a c such that for all n,

$$K((A \upharpoonright n)|(B \upharpoonright n + c)) = O(1).$$

The reducibility (i) is also called effective Lipschitz reducibility and This reducibility has been analyzed by Yu and Ding [105], Barmpalias and Lewis (e.g. [4]), and Raichev and Stephan (e.g. [79]). While I do not really have space to discuss these reducibilities in detail, I would like to point out that they do give nice insight into relative computability. We briefly consider sw. The idea of this reducibility is that if

$A \leq_{sw} B$, then there is an *efficient* way to convert the *bits* of B into those of A. The Kučera–Slaman Theorem says that all versions of Ω are the same in terms of their S-degrees. But we may ask whether there is a "bit" version of this result? Yu and Ding [105] established the following.

Theorem 4.2 (Yu and Ding [105]).

(i) *There is no sw-complete c.e. real.*

(ii) *There are two c.e. reals β_0 and β_1 so that there is no c.e. real α with $\beta_0 \leq_{sw} \alpha$ and $\beta_1 \leq_{sw} \alpha$.*

There are other assorted results and reducibilities. However, things are still in their infancy here. We will simply refer the reader to Downey [17], or Downey and Hirschfeldt [22] for the current situation.

We return to looking at the basic measures $\leq_C$ and $\leq_K$. The reader should note that these are not really reducibilities but simply transitive pre-orderings. (Though following tradition we will continue to refer to them as reducibilities.)

Theorem 4.3 (Yu, Ding, Downey [107]). *For $Q \in \{K, C\}$, $\{X : X \leq_Q Y\}$ has size $2^{\aleph_0}$ and has members of each degree, whenever Y is random.*

The replacement for this theorem is a measure-theoretical one:

Theorem 4.4 (Yu, Ding, Downey [107]). *For any real A, $\mu(\{B : B \leq_K A\}) = 0$. Hence there are uncountably many K degrees.*

We had hoped that there might be nice hierarchies related to levels of randomness. We will denote by $\Omega^{(m+1)}$ to be Ω relative to $\emptyset^{(m)}$. We might have hoped that $\Omega^{(2)}$ was K-above Ω, but that hope turns out to be forlorn.

Theorem 4.5 (Yu, Ding, Downey [107]). *For all c and $n < m$,*

$$(\exists^\infty k)\,[\,K(\Omega^{(n)} \upharpoonright k) < K(\Omega^{(m)} \upharpoonright k) - c].$$

For $n = 0$, $m = 1$ Theorem 4.5 was proven by Solovay [89], using totally different methods.

Miller and Yu have made really significant progress in our understanding here by introducing yet more measures of relative randomness. They are based around van Lambalgen's Theorem which states that for all A, B, *B n-random and A is B-n-random iff $A \oplus B$ is n-random.*

Definition 4.6 (Miller and Yu [69]). We say that $\alpha \leq_{vL} \beta$, α is *van Lambalgen*[11] reducible to β if for all $x \in 2^\omega$, $\alpha \oplus x$ is random implies $\beta \oplus x$ is random.

Miller and Yu's basic result were as follows.

[11] This is closely related to a relation introduced by Nies: He defined $A \leq_{LR} B$ if for all Z, Z is 1-B-random implies Z is 1-A-random. If A and B are both random then $A \leq_{LR} B$ iff $B \leq_{LR} A$.

Theorem 4.7 (Miller and Yu [69]). *For all random α, β,*

(i) *α n-random and $\alpha \leq_{vL} \beta$ implies β is n-random.*

(ii) *If $\alpha \oplus \beta$ is random then α and β have no upper bound in the vL-degrees.*

(iii) *If $\alpha \leq_T \beta$ and α is 1-random, then $\beta \leq_{vL} \alpha$.*

(iv) *There are random $\alpha \equiv_{vL} \beta$ of different Turing degrees.*

(v) *There are no maximal or minimal random vL-degrees, and no join.*

(vi) *If $\alpha \oplus \beta$ is random then $\alpha \oplus \beta <_{vL} \alpha, \beta$.*

(vii) *The Σ^0_1 theory of the vL-degrees is decidable.*

Miller and Yu show that $\Omega^{(n)}$ and $\Omega^{(m)}$ have no upper bound in the vL degrees for $n \neq m$. This improves the Yu, Ding, Downey (Theorem 4.5) result above. All of this is filters through an interesting relationship between $\leq_{vL}$ and $\leq_C, \leq_K$.

Lemma 4.8 (Miller and Yu [69]). *For random α, β,*

(i) *Suppose that $\alpha \leq_K \beta$. Then $\alpha \leq_{vL} \beta$.*

(ii) *Suppose that $\alpha \leq_C \beta$. Then $\alpha \leq_{vL} \beta$.*

We state the following for $\leq_K$ but they hold equally for $\leq_C$, as has been shown by Miller and Yu.

Corollary 4.9 (Miller and Yu [69]).

(i) *Suppose that $\alpha \leq_K \beta$, and α is n-random and β is random. Then β is n-random.*

(ii) *If $\alpha \oplus \beta$ is 1-random, then $\alpha|_K \beta$ and have no upper bound in the K-degrees.*

(iii) *For all $n \neq m$, the K-degrees of $\Omega^{(n)}$ and $\Omega^{(m)}$ have no upper bound.*

Miller and Yu have many other very interesting results on the K degrees of c.e. reals. For instance, they show that if $\alpha \oplus \beta$ is 1-random, then $\alpha|_K \alpha \oplus \beta$. Miller has proven the following.

Theorem 4.10 (Miller [67]).

(i) *If α, β are random, and $\alpha \equiv_K \beta$, then $\alpha' \equiv_{tt} \beta'$. As a consequence, every K-degree of a random real is countable.*

(ii) *If $\alpha \leq_K \beta$, and α is 3-random, then $\beta \leq_T \alpha \oplus \emptyset'$.*

Notice that (ii) implies that the cone of K-degrees above a 3-random is countable. On the other hand, Miller and Yu have constructed a 1-random whose K-upper cone is uncountable. The construction of an uncountable random K-degree uses their method of constructing K-comparable reals. Its proof uses the following clever lemma. The current proof of Theorem 4.11 is quite difficult.

Theorem 4.11 (Miller and Yu [70]). *Suppose that* $\sum_n 2^{-f(n)} < \infty$, *then there is a* 1*-random* Y *with*

$$K(Y \upharpoonright n) < n + f(n),$$

for almost all n.

To finish this section, we mention further evidence that randomness is a "lowness" notion. Miller has shown that if α is 3-random then its often useless as an oracle. We will call α *weakly-low for* K if $(\exists^\infty n)[K(n) \leq K^\alpha(n) + O(1)]$. Thus in a weakly-low real, the information in α is so useless that it cannot help to compress n. The following result echoes the theme articulated by Stephan that most random reals have little *usable* information in them.

Theorem 4.12 (Miller [67]).

(i) *If* α *is* 3*-random it is weakly-low for* K.

(ii) *If* α *is weakly-low for* K, *and random, then* α *is strongly Chaitin random in that*

$$(\exists^\infty n)\left[K(\alpha \upharpoonright n) \geq n + K(n) - O(1)\right].$$

5. Lowness and triviality

There have been some truly dramatic results in what has now become known as lowness and triviality. If Q is a measure of relative randomness then we can say that A is Q*-trivial* iff $A \leq_Q 1^\omega$. Thus using Q we cannot distinguish A from a computable set. We will say that A is Q*-low* if $Q^A(\sigma) = Q(\sigma) + O(1)$, for all σ. Thus, for instance A is K-low would mean that $K^A(\sigma) = K(\sigma) + O(1)$ for all σ.

We say that a set A is low for a randomness notion V iff the V-randoms relative to A remain the same. (One would usually expect that fewer sets would be random.) An apparently weaker notion is that of being low for V tests. That is, every if $\{U_i : i \in \mathbb{N}\}$ is a V^A test, then there is a V-test $\{\hat{U}_i : i \in \mathbb{N}\}$ such that $\cap_i U_i \subseteq \cap_i \hat{U}_i$. We remark that since there are universal Martin-Löf tests the test set notion and the lowness notion are the same.

5.1. The remarkable Martin-Löf case. There have been a series of amazing results in the case of 1-randomness. Historically, these results began with triviality. An old result of Loveland [57] shows that $Q(\alpha \upharpoonright n|n) = O(1)$ for all n, $(Q \in \{C, K\})$ iff α is computable. This result was generalized by Chaitin [11], who proved the following.

Theorem 5.1 (Chaitin [11]). α *is computable iff* $\alpha \leq_C 1^\omega$. (*That is, iff* α *is* C*-trivial.*)

I think this squares with our intuition that should α be indistinguishable from a computable string in terms of its initial segment complexity it should itself be

computable. Chaitin also noted that essentially the same proof shows that if α is K-trivial, the α is Δ^0_2 and hence computable from the halting problem. The breakthrough was again by Solovay.

Theorem 5.2 (Solovay [89]). *There are noncomputable α which are K-trivial.*

Solovay's argument was complex and mysterious. It turned out that the example α could even be chosen as a computably enumerable *set* (Calude and Coles [7], Downey, Hirschfeldt, Nies and Stephan [27], Kummer (unpubl.), An. A. Muchnik (unpubl.)). The paper [27] gave a very simple construction of a computably enumerable K-trivial set along the lines of the Dekker deficiency set. What is remarkable is that such K-trivial sets solve Post's problem.

Theorem 5.3 (Downey, Hirschfeldt, Nies and Stephan [27]). *Suppose that α is K-trivial. Then $\alpha <_T \emptyset'$.*

The method of proof of Theorem 5.3 uses what has become known as the "decanter method"(terminology of Nies) and is unfortunately very complicated, though it does *not* use the priority method. No easy proof of Theorem 5.3 is known.

It was noted that the short [27] proof constructing a K-trivial set strongly resembled and earlier construction of a computably enumerable set A which was low for Martin-Löf randomness by Kučera and Terwijn [47]. It was conjectured that perhaps these classes might have something to do with each other. In a ground breaking series of papers, Nies (and Hirschfeldt) proved some completely unexpected facts.

Theorem 5.4 (Nies (and Hirschfeldt for some), [75], [76]).

(a) *The following classes of reals coincide.*

  (i) *K-low.*
  (ii) *K-trivial.*
  (iii) *Low for Martin-Löf randomness.*

(b) *All the members A of this class $\mathcal{C}$ are* superlow *in that $A' \equiv_{wtt} \emptyset'$.*

(c) *The class $\mathcal{C}$ forms a natural Σ^0_3 ideal in the Turing degrees. There is a low$_2$ computably enumerable degree $\mathbf{a}$ such that if $\mathbf{c} \in \mathcal{C}$, the $\mathbf{c} < \mathbf{a}$.*

(d) *If A is a K-trivial real, then there is a computably enumerable set $\hat{A}$ with $A \leq_T \hat{A}$.*

The K-trivials form the only known natural nontrivial Σ^0_3 ideal in the (computably enumerable) Turing degrees. Item (c) in the above is a special case of a general unpublished Theorem of Nies that every Σ^0_3 ideal in the computably enumerable degrees is bounded by a low$_2$ computably enumerable degree. (A proof can be found in Downey and Hirschfeldt [22].) It is possible that there is a low (non-computably enumerable) degree $\mathbf{a}$ which bounds $\mathcal{C}$, and even possible that such a degree could be random. This problem seems hard.

Subsequently, other quite deep results have been proven. For instance, we have seen that if A is noncomputable then $\mu(\{X : A \leq_T X\}) = 0$, but since there are K-low reals, there must be reals A and randoms X such that X is A-random and $A \leq_T X$. In that case, we say that A is a *base of Martin-Löf randomness*.

Theorem 5.5 (Hirschfeldt, Nies, Stephan [37]). *A is K-trivial iff it is a base of Martin-Löf randomness.*

We remark that Slaman has used the class of K-trivials to solve a longstanding problem in computable model theory. As a final result in this area we mention some interesting results of Csima and Montalbán. These results are related to the enumeration of the K-trivials.

Theorem 5.6 (Chaitin [11], Zambella [108]). *There are only $O(2^d)$ members of $KT(d)$. They are all Δ^0_2.*

The reader might wonder with the nice computable bound how many K-trivial reals there are. Let $G(d) = |\{X : X \in KT(d)\}$. Then there is a crude estimate that $G(d) \leq_T \emptyset''$. This is the best upper bound known. In unpublished work, Downey, Miller and Yu have shown that $G(d) \not\leq_T \emptyset'$, using the fact that $\sum_d \frac{G(d)}{2^d}$ is convergent. This is all related to the Csima–Montalbán functions. We say that f is a *Csima–Montalbán function* if f is nondecreasing and

$$K(A \upharpoonright n) \leq K(n) + f(n) + O(1)$$

implies that $A \upharpoonright n$ is K-trivial. Such functions can be constructed from $\emptyset'' \oplus G$. We define f to be *weakly* Csima–Montalbán, if we weaken the hypothesis to be that $\liminf_n f(n) \to \infty$. Little is known here. It is not known if the arithmetical complexity of f depends upon the universal machine chosen. We remark that the original use of Csima–Montalbán functions was to construct a minimal pair of K-degrees: K-degrees $\mathbf{a}, \mathbf{b}$ such that $\mathbf{a} \wedge \mathbf{b} = \mathbf{0}$.

In other more recent work, Downey, Nies, Weber and Yu [29] have also looked at lowness for weak 2-randomness. Here it has been shown that such degrees do exist, and are all K-trivial. It is not known if the converse holds.

5.2. Other lowness and triviality. One thing which this work has brought (back) to the fore is the use of domination properties in classical computability. This was first recognized in the study of lowness for Schnorr randomness. Terwijn and Zambella [97] defined a degree $\mathbf{a}$ to be computably traceable iff there is a single computable function f such that for all functions $g \leq_T \mathbf{a}$, there is a computable collection of canonical finite sets $\{D_{p(x)} : x \in \mathbb{N}\}$, such that

(i) $|D_{p(x)}| < f(x)$, and

(ii) $g(x) \in D_{p(x)}$ for almost all x.

Being computably traceable is a strong form of being hyperimmune-free. Terwijn and Zambella showed that there are $2^{\aleph_0}$ many degrees that are hyperimmune-free yet not computably traceable. There are also $2^{\aleph_0}$ degrees that are computably traceable. The following theorem completely classifies the low for Schnorr random reals. Its proof is far from easy.

Theorem 5.7 (Terwijn and Zambella [97]). *A is low for Schnorr random null sets iff A is computably traceable.*

It is clear that if A is low for tests then A is low for Schnorr randoms. But the converse is not at all clear and had been an open question of Ambos-Spies and Kučera [3]. The question was finally solved by Kjos-Hanssen, Stephan, and Nies [40], using Bedregal and Nies [5]. Summarizing the results proven there, we have:

Theorem 5.8 (Kjos-Hanssen, Stephan, and Nies [40]). **a** *is low for Schnorr null sets iff* **a** *is low for Schnorr randomness.*

I remark in passing that I am not aware of any lowness notion that differs for null sets and for the randomness notion. In other work, Nies has examined lowness for polynomial time randomness, and lowness for computable randomness. For computable randomness, the answer is rather surprising.

Theorem 5.9 (Nies [76]). *Suppose that A is low for computable randomness. Then A is computable.*

Finally there has been a little work on triviality notions here. Recall that Downey and Griffiths [19] proved that A is Schnorr trivial iff for all *computable* machines M, $K_M(A \upharpoonright n) \geq n - O(1)$. This definition naturally allows us to define a reducibility notion.

Definition 5.10 (Downey and Griffiths [19]). We say that α is Schnorr reducible to β, $\alpha \leq_{Sch} \beta$, iff for all computable machines M, there is a computable machine $\widehat{M}$ such that $K_M(\beta \upharpoonright n) - O(1) > K_{\widehat{M}}(\alpha \upharpoonright n)$, for all n.

This definition allows us to say that a real α is *Schnorr trivial* iff $\alpha \leq_{Sch} 1^\omega$. Schnorr trivial reals behave quite differently than do Schnorr low reals and the K-trivials. Downey and Griffiths constructed a Schnorr trivial real and Downey, Griffiths and LaForte [20] showed that they can even be Turing complete, though they do not occur in every computably enumerable Turing degree. Subsequently, they have been investigated by Johanna Franklin [31]. Her results are summarized below.

Theorem 5.11 (Franklin [31]).

(i) *There is a perfect set of Schnorr trivials* (*and thus some are not* Δ^0_2).

(ii) *Every degree above* $\mathbf{0}'$ *contains a Schnorr trivial.*

(iii) *Every Schnorr low is Schnorr trivial.*

(iv) *The Schnorr lows are not closed under join.*

Finally, we mention that other lowness notions both in randomness and in other contexts have been analyzed. Yu [104] (also Miller and Greenberg (unpublished)) proved that there are no sets low for 1-genericity. Sets low for Kurtz randomness were first constructed by Downey, Griffiths and Reid [21]. They were shown there to be all hyperimmune-free and were implied by Schnorr lowness. Stephan and Yu [92] have shown that lowness for Kurtz randomness differs from lowness for Schnorr randomness and lowness for weak genericity. To wit, they have shown the following.

Theorem 5.12 (Stephan and Yu [92]).

(i) *Low for weakly generic is the same hyperimmune-free plus not of diagonally noncomputable degree.*

(ii) *There is a set of hyperimmune-free degree which is neither computably traceable nor diagonally noncomputable.*

(iii) *Low for weakly generic implies low for Kurtz random.*

(iv) *In particular, low for weakly generic and hence low for Kurtz randomness is not the same as Schnorr low.*

The topic of lowness for such concepts remains in its infancy, and promises fascinating results.

References

[1] Allender, E., Buhrman, H., Koucký, M., van Melkebeek, D., and Ronneburger, D., Power from Random Strings. In *FOCS* 2002*, IEEE* (2002), 669–678.

[2] Allender, E., Buhrman, H., and Koucký, M., What Can be Efficiently Reduced to the Kolmogorov-Random Strings? *Ann. Pure Appl. Logic* (2006) **138** (2006), 2–19.

[3] Ambos-Spies K., and Kučera, A., Randomness in computability theory. In *Computability Theory and its Applications* (ed. by P. A. Cholak, S. Lempp, M. Lerman and R. A. Shore) Contemporary Mathematics 257, Amer. Math. Soc., Providence, RI, 2000, 1–14.

[4] Barmpalias, G., and Lewis, A., Randomness and the Lipschitz degrees of computability. Submitted.

[5] Bedregal, B., and Nies, A., Lowness properties of reals and hyper-immunity. In *WOLLIC* 2003, *Electr. Notes Theor. Comput. Sci.* **84** (2003), Elsevier, 2003; http://www.cs.auckland.ac.nz/~nies/papers/benjanew.pdf.

[6] Calude, C., *Information Theory and Randomness. An Algorithmic Perspective*. EATCS Monographs on Theoretical Computer Science, Springer-Verlag, Berlin 1994; second revised edition 2002.

[7] Calude, C., and Coles, R., Program size complexity of initial segments and domination relation reducibility. In *Jewels are Forever* (ed. by J. Karhümaki, H. Mauer, G. Paŭn, G. Rozenberg), Springer-Verlag, Berlin 1999, 225–237.

[8] Calude, C., Coles, R., Hertling, P., Khoussainov, B., Degree-theoretic aspects of computably enumerable reals. In *Models and Computability* (ed. by S. B. Cooper and J. K. Truss), London Mathematical Soc. Lecture Note Ser. 259, Cambridge University Press, Cambridge 1999.

[9] Calude, C., Hertling, P., Khoussainov, B., Wang, Y., Recursively enumerable reals and Chaitin's Ω number. In *STACS '98*, Lecture Notes in Comput. Sci. 1373, Springer-Verlag, Berlin 1998, 596–606.

[10] Chaitin, G. J., A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.* **22** (1975), 329–340.

[11] Chaitin, G. J., Information-theoretic characterizations of recursive infinite strings. *Theoret. Comput. Sci.* **2** (1976), 45–48.

[12] Chaitin, G. J., Incompleteness theorems for random reals. *Adv. in Appl. Math* **8** (1987), 119–146.

[13] Chaitin, G. J., *Information, Randomness & Incompleteness*. 2nd edition, World Sci. Ser. Comput. Sci. 8, World Scientific, River Edge, NJ, 1990.

[14] de Leeuw, K., Moore, E. F., Shannon, C. E., and Shapiro, N., Computability by probabilistic machines. In *Automata studies*, Annals of Mathematics Studies 34, Princeton University Press, Princeton, N.J., 1956, 183–212.

[15] Downey, R., Some recent progress in algorithmic randomness. In *Proceedings of the* 29*th Annual Conference on Mathematical Foundations of Computer Science, Prague, August* 2004 (ed. by J. Fiala, V. Koubek and J. Kratochvil), Lecture Notes in Comput. Sci. 3153, Springer-Verlag, Berlin 2004, 42–81.

[16] Downey, R., Some Computability-Theoretical Aspects of Reals and Randomness. In *The Notre Dame Lectures* (ed. by P. Cholak), Lecture Notes in Logic 18, Association for Symbolic Logic, Urbana, IL, A K Peters, Ltd., Wellesley, MA, 2005, 97–146.

[17] Downey, R., Five lectures on algorithmic randomness. In *Proceedings of Computational Prospects of Inifinity* (edited by Chong et. al.), World Scientific, to appear.

[18] Downey, R., Ding, D., Tung S.-P., Qiu, Y.-H., Yasuugi, M., and Wu, G. (eds.). *Proceedings of the* 7*th and* 8*th Asian Logic Conferences*, Singapore University Press, Singapore; World Scientific Publishing Co., Inc., River Edge, NJ, 2003.

[19] Downey, R., and Griffiths, E., Schnorr randomness. *J. Symbolic Logic* **69** (2) (2004), 533–554.

[20] Downey, R., Griffiths, E., and LaForte, G., On Schnorr and computable randomness, martingales, and machines. *Math. Logic Quart.* **50** (2004), 613–627.

[21] Downey, R., Griffiths, E., and Reid, S., On Kurtz randomness. *Theoret. Comput. Sci.* **321** (2004), 249–270.

[22] Downey, R., and Hirschfeldt, D., *Algorithmic Randomness and Complexity*. Monographs in Computer Science, Springer-Verlag, to appear; preliminary version: www.mcs.vuw.ac.nz/~downey.

[23] Downey, R., Hirschfeldt, D., and LaForte, G., Randomness and reducibility. Extended abstract in *Mathematical Foundations of Computer Science,* 2001 (ed. by J. Sgall, A. Pultr, and P. Kolman), Lecture Notes in Comput. Sci. 2136 Springer-Verlag, Berlin 2001, 316–327; final version in *J. Comput. System Sci.* **68** (2004), 96–114.

[24] Downey, R., Hirschfeldt, D., and LaForte, G., Undecidability of Solovay degrees of c.e. reals. In preparation.

[25] Downey, R., Hirschfeldt, D., Miller, J., and Nies, A., Relativizing Chaitin's halting probability. *J. Math. Logic* **5** (2005), 167–192.

[26] Downey, R., Hirschfeldt, D., and Nies, A., Randomness, computability and density. *SIAM J. Comput.* **31** (2002), 1169–1183; extended abstract in *Proceedings of STACS* 2001 (ed. by A. Ferreira and H. Reichel), Lecture Notes in Comput. Sci. 2010, Springer-Verlag, Berlin 2001, 195–201.

[27] Downey, R., Hirschfeldt, D., Nies, A., and Stephan, F., Trivial reals. Extended abstract in *Computability and Complexity in Analysis* (ed. by V. Brattka, M. Schröder, K. Weihrauch), Electronic Notes in Theoretical Computer Science, FernUniversität Hagen, 294-6/2002, July 2002, 37–55; final version in [18], 103–131.

[28] Downey, R., Hirschfeldt, D., Nies, A., and Terwijn, S., Calibrating randomness. *Bull. Symbolic Logic*, to appear.

[29] Downey, R., Nies, A., Liang, Y., and Weber, R., Lowness and Π^0_2-Nullsets. In preparation.

[30] Downey, R., Merkle, W., and Reimann, J., Schnorr dimension. In *New Computational Paradigms: First Conference on Computability in Europe* (CiE 2005, Amsterdam, ed. by S. B. Cooper, B. Löwe, L. Torenvliet), Lecture Notes in Comput. Sci. 3526, Springer-Verlag, Berlin 2005, 96–105; final version to appear in *Mathematical Structures in Computer Science*.

[31] Franklin, J., Ph. D. Dissertation. University of California at Berkeley, in preparation.

[32] Gács, P., Exact Expressions for some Randomness Tests. *Z. Math. Logik Grundlag. Math.* **26** (1980), 385–394; short version in Lecture Notes in Comput. Sci. 67, Springer-Verlag, Berlin 1979, 124–131.

[33] Gács, P., Every set is reducible to a random one. *Inform. and Control* **70** (1986), 186–192.

[34] Gács, P., Uniform Test of Algorithmic Randomness Over a General Space. Online manuscript.

[35] Gaifmann, H., and Snir, M., Probabilities over rich languages. *J. Symbolic Logic* **47** (1982), 495–548.

[36] Hausdorff, F., Dimension und äußeres Maß. *Math. Ann.* **79** (1919) 157–179.

[37] Hirschfeldt, D., Nies, A., and Stephan, F., Using random sets as oracles. To appear.

[38] Kautz, S., Degrees of Random Sets. Ph.D. Thesis, Cornell University, 1991.

[39] Kjos-Hanssen, B., Merkle, W. , and Stephan, F., Kolmogorov complexity and the Recursion Theorem. To appear.

[40] Kjos-Hanssen, B., Stephan, F., and Nies, A., On a question of Ambos-Spies and Kučera. To appear.

[41] Kolmogorov, A. N., Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* **1** (1965), 3–11; English translation *Internat. J. Comput. Math.* **2** (1968), 157–168.

[42] Kučera, A., Measure, Π^0_1 classes, and complete extensions of PA. In *Recursion theory week* (Oberwolfach 1984), Lecture Notes in Math. 1141, Springer-Verlag, Berlin 1985, 245–259.

[43] Kučera, A., An alternative, priority-free solution to Post's Problem. in *Mathematical Foundations of Computer Science* (ed. by J. Gruska, B. Rovan, and J. Wiederman), Lecture Notes in Comput. Sci. 233, Springer-Verlag, Berlin 1986, 493–500.

[44] Kučera, A., On the use of diagonally nonrecursive functions. In *Logic Colloquium* '87, Granada, 1987), Stud. Logic Found. Math. 129, North-Holland, Amsterdam 1989, 219–239.

[45] Kučera, A., Randomness and generalizations of fixed point free functions. In *Recursion theory week* (Oberwolfach 1989, ed. by K. Ambos-Spies, G. H. Müller and G. E. Sacks), Lecture Notes in Math. 1432, Springer-Verlag, Berlin 1990, 245–254.

[46] Kučera, A., and Slaman, T., Randomness and recursive enumerability. *SIAM J. Comput.* **31** (2001), 199–211.

[47] Kučera, A., and Terwijn, S., Lowness for the class of random sets. *J. Symbolic Logic* **64** (1999), 1396–1402.

[48] Kummer, M., Kolmogorov complexity and instance complexity of recursively enumerable sets. *SIAM J. Comput.* **25** (1996), 1123–1143.

[49] Kummer, M., On the complexity of random strings. Extended abstract in *STACS* '96, Lecture Notes in Comput. Sci. 1046, Springer-Verlag, Berlin 1996, 25–36.

[50] Kurtz, S., Randomness and Genericity in the Degrees of Unsolvability. Ph. D. Thesis, University of Illinois at Urbana, 1981.

[51] Levin, L., Some Theorems on the Algorithmic Approach to Probability Theory and Information Theory. Dissertation in Mathematics, Moscow, 1971.

[52] Levin, L., On the notion of a random sequence. *Soviet Math. Dokl.* **14** (1973) 1413–1416.

[53] Levin, L., Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problemy Peredači Informacii* **10** (3) (1974), 30–35.

[54] Levin, L., Measures of complexity of finite objects (axiomatic description). *Soviet Math. Dokl.* **17** (1976), 522–526.

[55] Levy, P., *Theorie de l'Addition des Variables Aléatoires*. Monographies des probabilités 1, Gauthier-Villars, Paris 1937.

[56] Li, M., and Vitanyi, P., *Kolmogorov Complexity and its Applications*. Texts Monogr. Comput. Sci., Springer-Verlag, New York 1993.

[57] Loveland, D., A variant of the Kolmogorov concept of complexity. *Inform. and Control* **15** (1969), 510–526.

[58] Lutz, J. H., The dimensions of individual strings and sequences. *Inform. and Comput.* **187** (2003), 49–79; preliminary version: Gales and the constructive dimension of individual sequences, in *Automata, Languages, and Programming* (ed. by U. Montanari, J. D. P. Rolim, E. Welzl), Lecture Notes in Comput. Sci. 1853, Springer, Berlin 2000, 902–913.

[59] Lutz, J., Effective fractal dimensions. *Math. Logic Quart.* **51** (2005), 62–72.

[60] Martin-Löf, P., The definition of random sequences. *Inform. and Control* **9** (1966), 602–619.

[61] Martin-Löf, P., Complexity oscillations in infinite binary sequences. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **19** (1971), 225–230.

[62] Merkle W., The complexity of stochastic sequences. Preliminary version in CCC2003, final version to appear.

[63] Merkle W., The Kolmogorov-Loveland stochastic sequences are not closed under selecting subsequences. *J. Symbolic Logic* **68** (2003), 1362–1376.

[64] Merkle W., and Mihailovic, N., On the construction of effective random sets. In *Mathematical foundations of computer science* 2002, Lecture Notes in Comput. Sci. 2420, Springer-Verlag, Berlin 2002, 568–580.

[65] Merkle, W., Miller, J., Nies, A., Reimann, J., and Stephan, F., Kolmogorov-Loveland randomness and stochasticity. *Ann. Pure Appl. Logic* **138** (2006), 183–210.

[66] J. S. Miller, Kolmogorov random reals are 2-random. *J. Symbolic Logic* **69** (2004), 907–913.

[67] Miller, J., The K-degrees, low for K degrees, and weakly low for K oracles. In preparation.

[68] Miller, J., Contrasting plain and prefix-free Kolmogorov complexity. In preparation.

[69] Miller, J., and Yu, L., On initial segment complexity and degrees of randomness. *Trans. Amer. Math. Soc*., to appear.

[70] Miller, J., and Yu, L., Oscillation in the initial segment complexity of random reals. In preparation.

[71] Muchnik, An. A., and Positelsky, S. P., Kolmogorov entropy in the context of computability theory. *Theor. Comput. Sci.* **271** (2002), 15–35.

[72] Muchnik, A. A., Semenov, A., and Uspensky, V., Mathematical metaphysics of randomness. *Theor. Comput. Sci.* **207** (1998), 263–317.

[73] Miller, W., and Martin, D. A., The degree of hyperimmune sets. *Z. Math. Logik Grundlagen Math.* **14** (1968), 159–166.

[74] Ng, K. M., Stephan, F., and Wu, G., The Degrees of Weakly Computable Reals. In preparation.

[75] Nies, A., Reals which compute little. In *Proceedings of CL* 2002, to appear.

[76] Nies, A., Lowness properties and randomness. *Adv. Math.* **197** (2005), 274–305.

[77] Nies, A., *Computability and Randomness*. Monograph in preparation.

[78] Nies, A., Stephan, F., and Terwijn, S. A., Randomness, relativization, and Turing degrees. *J. Symbolic Logic* **70** (2005), 515–535.

[79] Raichev, A., Ph.D. Thesis. University of Wisconsin-Madison, in progress.

[80] Reimann, J., Computability and Dimension. PhD Thesis, University of Heidelberg, 2004

[81] Sacks, G. E., *Degrees of Unsolvability*. Princeton University Press, Princeton, N.J., 1963.

[82] Schnorr, C. P., A unified approach to the definition of random sequences. *Math. Systems Theory* **5** (1971), 246–258.

[83] Schnorr, C. P., *Zufälligkeit und Wahrscheinlichkeit*. Lecture Notes in Math. 218, Springer-Verlag, Berlin, New York 1971.

[84] Schnorr, C. P., Process complexity and effective random tests. *J. Comput. System Sci.* **7** (1973), 376–388.

[85] Scott, D., Algebras of sets binumerable in complete extensions of arithmetic. *Proc. Symp. Pure Appl. Math* **5** (1962), 357–366.

[86] Soare, R., *Recursively enumerable sets and degrees*. Perspect. Math. Logic, Springer-Verlag, Berlin 1987.

[87] Soare, R., Computability Theory and Differential Geometry. *Bull. Symbolic Logic* **10** (2004), 457–486.

[88] Solomonoff, R., A formal theory of inductive inference. I. *Inform. and Control* **7** (1964), 1–22; A formal theory of inductive inference. II. Ibid. 224–254.

[89] Solovay, R., Draft of paper (or series of papers) on Chaitin's work. Unpublished notes, May, 1975, 215 pages.

[90] Staiger, L., Kolmogorov complexity and Hausdorff dimension. *Inform. and Comput*. **103** (1993), 159–194.

[91] Stephan, F., Martin-Löf random sets and PA-complete sets. Forschungsberichte Mathematische Logik 58, Mathematisches Institut, Universität Heidelberg, Heidelberg, 2002.

[92] Stephan, F., and Liang, Y., Lowness for weakly 1-generic and Kurtz random. In preparation.

[93] Stillwell, J., Decidability of "almost all" theory of degrees. *J. Symbolic Logic* **37** (1972), 501–506.

[94] Tadaki, K., A generalization of Chaitin's halting probability Ω and halting self-similar sets. *Hokkaido Math. J.* **32** (2002), 219–253.

[95] Terwijn, S. Computability and Measure. Ph. D. Thesis, University of Amsterdam, 1998.

[96] Terwijn, S. A., *Complexity and Randomness*. Notes for a course given at the University of Auckland, March 2003, published as research report CDMTCS-212, University of Auckland.

[97] Terwijn, S. A., and Zambella, D., Computational randomness and lowness. *J. Symbolic Logic* **66** (2001) 1199–1205.

[98] Uspensky, V., Semenov, A., and Shen, A. Kh., Can an Individual Sequence of Zeros and Ones be Random? *Russian Math. Surveys* **45** (1990), 121–189.

[99] van Lambalgen, M., Random Sequences. Ph. D. Dissertation, University of Amsterdam, 1987.

[100] van Lambalgen, M., The axiomatization of randomness. *J. Symbolic Logic* **55** (1990), 1143–1167.

[101] von Mises, R., Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Z.* **5** (1919), 52–99.

[102] Ville, J., *Étude critique de la notion du collectif*. Monographies des probabilités 3, Gauthier-Villars, Paris 1939.

[103] Wang, Y., Randomness and Complexity. Ph. D. Dissertation, University of Heidelberg, 1996.

[104] Yu, Liang, Lowness for genericity. *Arch. Math. Logic* **45** (2006), 233–238.

[105] Yu, L., and Ding, D., There is no sw-complete c.e. real. *J. Symbolic Logic* **69** (2004), 1163–1170.

[106] Yu, L., and Ding, D., There are $2^{\aleph_0}$ many H-degrees in the random reals. *Proc. Amer. Math. Soc.* **132** (2004), 2461–2464

[107] Yu, L., Ding, D., and Downey, R., The Kolmogorov complexity of random reals. *Ann. Pure Appl. Logic* **129** (1–3) (2004), 163–180.

[108] Zambella, D., On sequences with simple initial segments. ILLC technical report, ML-1990-05, University of Amsterdam, 1990.

[109] Zvonkin A. K., and L.A. Levin, The complexity of finite objects and the development of concepts of information and randomness by the theory of algorithms. *Russian Math. Surveys* **25** (6) (1970), 83–124.

School of Mathematical and Computing Sciences, Victoria University, PO Box 600, Wellington, New Zealand

E-mail: rod.downey@vuw.ac.nz

URL: http://www.mcs.vuw.ac.nz/~downey

Determinacy and large cardinals

Itay Neeman*

Abstract. The principle of determinacy has been crucial to the study of definable sets of real numbers. This paper surveys some of the uses of determinacy, concentrating specifically on the connection between determinacy and large cardinals, and takes this connection further, to the level of games of length ω_1.

Mathematics Subject Classification (2000). 03E55; 03E60; 03E45; 03E15.

Keywords. Determinacy, iteration trees, large cardinals, long games, Woodin cardinals.

1. Determinacy

Let ω^ω denote the set of infinite sequences of natural numbers. For $A \subset \omega^\omega$ let $G_\omega(A)$ denote the length ω game with payoff A. The format of $G_\omega(A)$ is displayed in Diagram 1. Two players, denoted I and II, alternate playing natural numbers forming together a sequence $x = \langle x(n) \mid n < \omega \rangle$ in ω^ω called a *run* of the game. The run is won by player I if $x \in A$, and otherwise the run is won by player II.

I	$x(0)$		$x(2)$		.....	
II		$x(1)$		$x(3)$		.....

Diagram 1. The game $G_\omega(A)$.

A game is *determined* if one of the players has a winning strategy. The set A is determined if $G_\omega(A)$ is determined. For $\Gamma \subset \mathcal{P}(\omega^\omega)$, $\det(\Gamma)$ denotes the statement that all sets in Γ are determined.

Using the axiom of choice, or more specifically using a wellordering of the reals, it is easy to construct a non-determined set A. $\det(\mathcal{P}(\omega^\omega))$ is therefore false. On the other hand it has become clear through research over the years that $\det(\Gamma)$ is true if all the sets in Γ are definable by some concrete means. Moreover $\det(\Gamma)$, taken as an axiom, gives rise to a rich structure theory that establishes a hierarchy of complexity on the sets in Γ, and completely answers all natural questions about the sets in each level of the hierarchy. Determinacy is therefore accepted as the natural hypothesis in the study of definable subsets of ω^ω.

*This material is based upon work supported by the National Science Foundation under Grant No. DMS-0094174.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Definability takes increasingly liberal meanings as one progresses higher in the hierarchy of complexity. At the lower levels it is very concrete. Let $\omega^{<\omega}$ denote the set of finite sequences of natural numbers. For $s \in \omega^{<\omega}$ let $N_s = \{x \in \omega^\omega \mid x \text{ extends } s\}$. The sets N_s ($s \in \omega^{<\omega}$) are the *basic open* neighborhoods in ω^ω. $A \subset \omega^\omega$ is *open* if it is a union of basic open neighborhoods.

ω^ω with the topology defined above is isomorphic to the irrational numbers. Following standard abuse of notation in descriptive set theory we use $\mathbb{R}$ to denote ω^ω, and refer to its elements as *reals*.

A set is *Borel* if it can be obtained from open sets using repeated applications of complementations and countable unions. The *projection* of a set $B \subset \mathbb{R} \times \mathbb{R}$ is the set $\{x \in \mathbb{R} \mid (\exists y)\langle x, y\rangle \in B\}$. A set is *analytic* if it is the projection of the complement of an open set. A set is *projective* if it can be obtained from open sets using repeated applications of complementations and projections. Analyzing the logical complexity of these definitions and using diagonal arguments one can establish that {Borel sets} $\subsetneq$ {analytic sets} $\subsetneq$ {projective sets}, so that these classes form a proper hierarchy.

Theorem 1.1 (Gale–Stewart [4], 1953). *All open sets are determined.*

Theorem 1.2 (Martin [20], 1975). *All Borel sets are determined.*

Theorem 1.3 (Martin [19], 1970). *All analytic sets are determined.*

Theorem 1.4 (Martin–Steel [22], 1985). *All projective sets are determined.*

Theorems 1.1 and 1.2 are theorems of ZFC, the basic system of axioms for set theory and mathematics. Theorems 1.3, 1.4, and 1.5 (below) have additional stronger assumptions known as large cardinal axioms, which are not listed here but will be discussed in Section 2.[1]

Recall that $\mathrm{L}(\mathbb{R})$ is the smallest model of set theory which contains all the reals and all the ordinals. It is obtained as the union $\bigcup_{\alpha\in\mathrm{ON}} \mathrm{L}_\alpha(\mathbb{R})$ of the hierarchy defined by the conditions: $\mathrm{L}_0(\mathbb{R}) = \mathbb{R}$; for limit ordinals λ, $\mathrm{L}_\lambda(\mathbb{R}) = \bigcup_{\alpha<\lambda} \mathrm{L}_\alpha(\mathbb{R})$; and for each ordinal α, $\mathrm{L}_{\alpha+1}(\mathbb{R})$ consists of the sets in $\mathrm{L}_\alpha(\mathbb{R})$, and of all subsets of $\mathrm{L}_\alpha(\mathbb{R})$ which are definable over $\mathrm{L}_\alpha(\mathbb{R})$ by first order formulae with parameters. The third condition is the crucial one, placing a definability requirement on the sets that make it into $\mathrm{L}(\mathbb{R})$. $\mathrm{L}(\mathbb{R})$ is constructed through a transfinite sequence of applications of this condition. Notice that the projective sets are subsumed already into $\mathrm{L}_1(\mathbb{R})$, the first stage of this transfinite sequence.

Theorem 1.5 (Woodin [40], 1985). *All sets of reals in* $\mathrm{L}(\mathbb{R})$ *are determined.*

Theorems 1.1 through 1.5 establish determinacy for sets of varying levels of definability, starting from open sets which are very directly definable from real numbers,

[1] The determinacy of Borel sets of course follows from the determinacy of analytic sets. The new element in Theorem 1.2 is a proof of Borel determinacy from the axioms of ZFC, without using large cardinals.

continuing with the projective sets, which are definable from open sets using existential quantifications and negations, and ending with all sets in $\mathrm{L}(\mathbb{R})$. More is possible, as we shall see in Section 3. The remainder of this section is devoted to consequences of determinacy.

Let Γ be an adequate pointclass (that is a collection of subsets of ω^ω, closed under some basic operations, see Moschovakis [27]). The first results derived from determinacy concerned regularity properties, such as Lebesgue measurability, the Baire property, and the perfect set property. All these properties fail outside the realm of determinacy; counterexamples to each of them can be constructed easily using a wellordering of the reals. Determinacy serves as an intermediary in establishing these properties for definable sets.

Theorem 1.6 (Banach, Oxtoby [37], 1957). *Assume* $\det(\Gamma)$. *Let* $A \in \Gamma$. *Then* A *has the property of Baire* (*meaning that* A *is either meager, or comeager on a basic open neighborhood*).

Theorem 1.7 (Mycielski–Swierczkowski [29], 1964). *Assume* $\det(\Gamma)$. *Then all sets in* Γ *are Lebesgue measurable.*

Theorem 1.8 (Davis [3], 1964). *Assume* $\det(\Gamma)$. *Let* $A \in \Gamma$. *Then either* A *is countable, or else* A *contains a perfect set.*

More importantly, determinacy was seen to imply various structural properties of classes of sets within its realm. For a pointclass Γ let $\neg\Gamma$ denote the pointclass consisting of complements of sets in Γ, and let $\exists\Gamma$ denote the pointclass consisting of projections of sets in Γ. Recall that $\mathbf{\Sigma}^1_1$ is the pointclass of analytic sets, $\mathbf{\Pi}^1_n = \neg\mathbf{\Sigma}^1_n$, and $\mathbf{\Sigma}^1_{n+1} = \exists\mathbf{\Pi}^1_n$. $\mathbf{\Delta}^1_n$ is the pointclass consisting of sets which are both $\mathbf{\Sigma}^1_n$ and $\mathbf{\Pi}^1_n$. Each $\mathbf{\Sigma}^1_n$ set A (similarly $\mathbf{\Pi}^1_n$) is definable through a string of quantifiers from an open set. The open set itself, call it D, is definable from a real number, coding the set $\{s \in \omega^{<\omega} \mid N_s \subset D\}$. A is *lightface* Σ^1_n (similarly Π^1_n) if the underlying real that defines it is recursive, that is computable by a Turing machine.

The boldface pointclasses were studied by analysts in the early 20th century. Recall for example the following theorem of Kuratowski [16]: the intersection of any two $\mathbf{\Sigma}^1_1$ (analytic) sets $A, B \subset \mathbb{R}$ can be presented as the intersection of two $\mathbf{\Sigma}^1_1$ sets $A' \supset A$ and $B' \supset B$, such that $A' \cup B' = \mathbb{R}$. This is today recast as a theorem about the pointclass $\mathbf{\Pi}^1_1$. A pointclass Γ is said to have the *reduction property* if for any two sets $A, B \subset \mathbb{R}$ in Γ there are sets $A' \subset A$ and $B' \subset B$, both in Γ, so that $A' \cup B' = A \cup B$ and $A' \cap B' = \emptyset$. Kuratowski's theorem establishes that $\mathbf{\Pi}^1_1$ has the reduction property. Kuratowski also showed that $\mathbf{\Sigma}^1_2$ has the reduction property. This was as far up along the projective hierarchy as one could get in those days. The basic axioms of set theory, without the addition of determinacy or large cardinals, do not decide questions such as the reduction property for projective pointclasses above $\mathbf{\Sigma}^1_2$.

In 1967 Blackwell [2] used the determinacy of open games, Theorem 1.1, to give a new proof of Kuratowski's reduction theorem. Inspired by his proof, Martin [18]

and Addison–Moschovakis [1] proved that $\mathbf{\Pi}^1_3$ has the reduction property, assuming $\det(\mathbf{\Delta}^1_2)$.

The reduction property is a consequence of a stronger property known as the prewellordering property. Martin and Addison–Moschovakis obtained this stronger property, and in fact propagated it along the odd levels of the projective hierarchy, using determinacy.

A *prewellorder* on $A \subset \mathbb{R}$ is a relation $\preceq$ on A which is transitive, reflexive, and wellfounded. The prewellorder $\preceq$ induces an equivalence relation $\sim$ on A ($x \sim y$ iff $x \preceq y \wedge y \preceq x$), and gives rise to a wellorder of $A/\sim$. $\preceq$ is said to belongs to a pointclass Γ if there are two relations Y and N, in Γ and $\neg\Gamma$ respectively, so that for every $y \in A$, $\{x \mid x \preceq y\} = \{x \mid \langle x, y\rangle \in Y\} = \{x \mid \langle x, y\rangle \in N\}$. Γ has the *prewellordering property* just in case that every set $A \in \Gamma$ admits a prewellorder in Γ.

Theorem 1.9 (Martin [18], Addison–Moschovakis [1], 1968). *Assume projective determinacy. Then the projective pointclasses with the prewellordering* (*similarly reduction*) *property are* $\mathbf{\Pi}^1_1$, $\mathbf{\Sigma}^1_2$, $\mathbf{\Pi}^1_3$, $\mathbf{\Sigma}^1_4$, $\mathbf{\Pi}^1_5$, ...

Remark 1.10. For $B \subset \mathbb{R} \times \mathbb{R}$ and $x \in \mathbb{R}$ let B_x denote $\{y \mid \langle x, y\rangle \in B\}$. Recall that $\circlearrowright B$ is the set $\{x \in \mathbb{R} \mid$ player I has a winning strategy in $G_\omega(B_x)\}$. It is common to write $(\circlearrowright y)B(x, y)$, or $(\circlearrowright y)\langle x, y\rangle \in B$, for the statement $x \in \circlearrowright B$. This notation is similar to the notation used for the quantifiers $(\forall y)$ and $(\exists y)$, and $(\circlearrowright y)$ too is viewed as a quantifier, giving precise meaning to the chain $(\exists y(0))(\forall y(1))(\exists(y(2)) \cdots\cdots$ of quantifiers over ω. For a pointclass Γ let $\circlearrowright\Gamma = \{\circlearrowright B \mid B \in \Gamma\}$. It is easy to check that $\circlearrowright\mathbf{\Pi}^1_n = \mathbf{\Sigma}^1_{n+1}$ and (using determinacy) $\circlearrowright\mathbf{\Sigma}^1_n = \mathbf{\Pi}^1_{n+1}$. Theorem 1.9 therefore states that the pointclasses $\circlearrowright^{(n)}\mathbf{\Pi}^1_1$, $n < \omega$, all have the reduction and prewellordering properties.

Theorem 1.9 helped establish the use of determinacy as a hypothesis in the study of definable sets of reals. In particular it became standard to study $\mathrm{L}(\mathbb{R})$ using the relativization to $\mathrm{L}(\mathbb{R})$ of the assumption that *all* sets of reals are determined, known as the axiom of determinacy (AD) and initially advanced by Mycielski–Steinhaus [28]. The use of this assumption in $\mathrm{L}(\mathbb{R})$ is justified in retrospect by Theorem 1.5.

There has been a wealth of results on sets of reals, on structural properties of pointclasses, and on $\mathrm{L}(\mathbb{R})$, assuming determinacy. Only a couple of results, the ones which are directly relevant to this paper, are listed below. A more complete account can be found in Moschovakis [27] and in the Cabal volumes [13], [10], [11], [12].

Recall that the symbol $\boldsymbol{\delta}$ is used to denote the supremum of the ordertypes of $\mathbf{\Delta}$ prewellorders on $\mathbf{\Delta}$ sets.

Theorem 1.11. *Assume* AD. *Then* $\boldsymbol{\delta}^1_1$ *is equal to* ω_1, $\boldsymbol{\delta}^1_2$ *is equal to* ω_2 (*Martin*), *and* $\boldsymbol{\delta}^1_3$ *is equal to* $\omega_{\omega+1}$ (*Martin*). *Much more is known, see Kechris* [9] *and Jackson* [7].

The values of the ordinals $\boldsymbol{\delta}^1_1$, $\boldsymbol{\delta}^1_2$, etc. are absolute between $\mathrm{L}(\mathbb{R})$ and the true universe V. Theorem 1.11 therefore implies that $\boldsymbol{\delta}^1_1 = (\omega_1)^{\mathrm{L}(\mathbb{R})}$, $\boldsymbol{\delta}^1_2 = (\omega_2)^{\mathrm{L}(\mathbb{R})}$, and $\boldsymbol{\delta}^1_3 = (\omega_{\omega+1})^{\mathrm{L}(\mathbb{R})}$. ω_1 is absolute between $\mathrm{L}(\mathbb{R})$ and V, so $\boldsymbol{\delta}^1_1 = \omega_1$. But other

cardinals need not be absolute. Theorem 1.11 by itself therefore does not provide information on the cardinalities of $\boldsymbol{\delta}^1_2$ and $\boldsymbol{\delta}^1_3$.

Theorem 1.12 (Steel–Van Wesep [38], Woodin [39]). *Assume* $\mathsf{AD}^{\mathrm{L}(\mathbb{R})}$. *Then it is consistent* (*with* $\mathsf{AD}^{\mathrm{L}(\mathbb{R})}$ *and the axiom of choice*) *that* $(\omega_2)^{\mathrm{L}(\mathbb{R})} = \omega_2$, *and hence that* $\boldsymbol{\delta}^1_2 = \omega_2$.

Note that the statement that $\boldsymbol{\delta}^1_2 = \omega_2$ implies a strong failure of the continuum hypothesis: not only must the continuum have size at least ω_2, but this must be witnessed by $\boldsymbol{\Delta}^1_2$ prewellorders.

2. Large cardinals

An embedding $\pi : P \to M$ is *elementary* just in case that it preserves truth, meaning that $\varphi[x_1, \ldots, x_k]$ holds in P iff $\varphi[\pi(x_1), \ldots, \pi(x_k)]$ holds in M, for all formulae φ and all $x_1, \ldots, x_k \in P$. Large cardinal axioms state the existence of elementary embeddings of the universe. For example, a cardinal κ is *measurable* if it is the critical point of an elementary embedding $\pi : \mathrm{V} \to M \subset \mathrm{V}$. The axiom "there exists a measurable cardinal" thus asserts the existence of a non-trivial elementary embedding acting on the entire universe.

An elementary embedding $\pi : \mathrm{V} \to M$ is λ-*strong* if M and V agree to λ, that is if M and V have the same bounded subsets of λ, and *superstrong* if M and V agree to $\pi(\mathrm{crit}(\pi))$. $\pi : \mathrm{V} \to M$ is λ-*strong with respect to* H if it is λ-strong and $\pi(H) \cap \lambda = H \cap \lambda$. κ is λ-strong if it is the critical point of a λ-strong embedding, and similarly for superstrength and strength with respect to H. κ is $<\delta$-strong with respect to H if it is λ-strong with respect to H for each $\lambda < \delta$. Finally, and most importantly, δ is a *Woodin cardinal* if for every $H \subset \delta$ there is $\kappa < \delta$ which is $<\delta$-strong with respect to H. In the hierarchy of large cardinal axioms, the existence of Woodin cardinals lies above the existence of measurable cardinals, but well below the existence of superstrong cardinals.

Let $\pi : \mathrm{V} \to M$ be elementary. Let $\kappa = \mathrm{crit}(\pi)$ and let $\lambda < \pi(\kappa)$. The (κ, λ)-*extender* induced by π is the function $E : \mathcal{P}([\kappa]^{<\omega}) \to \mathcal{P}([\lambda]^{<\omega})$ defined by $E(A) = \pi(A) \cap [\lambda]^{<\omega}$. The extender E codes enough of the embedding π to reconstruct an embedding $\sigma : \mathrm{V} \to N$ with the property that $\sigma(A) \cap [\lambda]^{<\omega} = \pi(A) \cap [\lambda]^{<\omega}$ for all $A \subset [\kappa]^{<\omega}$. For sufficiently closed λ this is enough that the λ-strength of π implies the λ-strength of σ, and similarly for strength with respect to H. Thus, the existence of strong embeddings is equivalent to the existence of strong extenders, and the property of being a Woodin cardinal can be recast as a statement about the existence of certain extenders. (The point here is that extenders are sets, while embeddings are classes.)

The embedding $\sigma : \mathrm{V} \to N$ is obtained from the extender E using an ultrapower construction. Very briefly, N is the model $(H/\!\sim; R)$ where $H = \{\langle f, a\rangle \mid a \in [\lambda]^{<\omega}$ and $f : [\kappa]^{\mathrm{lh}(a)} \to \mathrm{V}\}$, $\langle f, a\rangle \sim \langle g, b\rangle$ iff $a^\frown b \in E(\{x^\frown y \mid f(x) = g(y)\})$, and $[f, a]\ R\ [g, b]$ iff $a^\frown b \in E(\{x^\frown y \mid f(x) \in g(y)\})$. The embedding σ is defined

by the conditions $\sigma(x) = [c_x, \emptyset]$ where $c_x : [\kappa]^0 \to \mathrm{V}$ is the function taking constant value x. The model N is called the *ultrapower* of V by E, denoted $\mathrm{Ult}(\mathrm{V}, E)$, and σ is the *ultrapower embedding*.

An extender E can also be derived from an embedding $\pi : Q \to M$ for $Q \neq \mathrm{V}$. The result is an *extender over* Q. In the other direction, the ultrapower of a model Q by an extender E with critical point κ can be defined so long as $(\mathcal{P}([\kappa]^{<\omega}))^Q = \mathrm{dom}(E)$, simply by adding the restriction $f \in Q$ to the definition of H above. The resulting ultrapower is denoted $\mathrm{Ult}(Q, E)$.

The process of taking ultrapowers can be iterated, and such iterations are crucial to the study of large cardinals. Their first use appeared in Kunen [15]. Kunen worked with measurable cardinals. The associated extenders can only give rise to linear iteration, and this has become the norm until the work of Martin–Steel [23], who introduced the general format of an iteration tree. This general format, which allows non-linearity, is both necessary to the study of Woodin cardinals, and non-trivial in their presence.

A *tree order* on an ordinal α is an order T so that: T is a suborder of $< \restriction \alpha$; for each $\eta < \alpha$, the set $\{\xi \mid \xi \ T \ \eta\}$ is linearly ordered by T; for each ξ so that $\xi + 1 < \alpha$, the ordinal $\xi + 1$ is a successor in T; and for each limit ordinal $\gamma < \alpha$, the set $\{\xi \mid \xi \ T \ \gamma\}$ is cofinal in γ. An *iteration tree* $\mathcal{T}$ of length α on a model M consists of a tree order T on α, and sequences $\langle M_\xi, j_{\zeta,\xi} \mid \zeta \ T \ \xi < \alpha\rangle$ and $\langle E_\xi \mid \xi + 1 < \alpha\rangle$ satisfying the following conditions:

1. $M_0 = M$.
2. For each ξ so that $\xi + 1 < \alpha$, E_ξ is an extender of M_ξ.
3. $M_{\xi+1} = \mathrm{Ult}(M_\zeta, E_\xi)$ and $j_{\zeta,\xi+1} : M_\zeta \to M_{\xi+1}$ is the ultrapower embedding, where ζ is the T-predecessor of $\xi + 1$. It is implicit in this condition that $\mathcal{P}([\mathrm{crit}(E_\xi)]^{<\omega})$ must be the same in M_ζ and M_ξ, so that the ultrapower makes sense.
4. For limit $\lambda < \alpha$, M_λ is the direct limit of the system $\langle M_\zeta, j_{\zeta,\xi} \mid \zeta \ T \ \xi \ T \ \lambda\rangle$, and $j_{\zeta,\lambda} : M_\zeta \to M_\lambda$ for $\zeta \ T \ \lambda$ are the direct limit embeddings.
5. The remaining embeddings $j_{\zeta,\xi}$ for $\zeta \ T \ \xi < \alpha$ are obtained through composition.

An iteration tree is *linear* if for every ξ, the T-predecessor of $\xi + 1$ is ξ.

A *branch* through an iteration tree $\mathcal{T}$ is a set b which is linearly ordered by T. The branch is *cofinal* if $\sup(b) = \mathrm{lh}(\mathcal{T})$. The branch is maximal if either $\sup(b) = \mathrm{lh}(\mathcal{T})$ or else $b \neq \{\xi \mid \xi \ T \ \sup(b)\}$. The *direct limit* along b, denoted $M_b^{\mathcal{T}}$ or simply M_b, is the direct limit of the system $\langle M_\xi, j_{\zeta,\xi} \mid \zeta \ T \ \xi \in b\rangle$. $i_b^{\mathcal{T}} : M \to M_b$ is the direct limit embedding of this system. The branch b is called *wellfounded* just in case that the model M_b is wellfounded.

Theorem 2.1 (Martin–Steel [23]). *Let M be a countable elementary substructure of a rank initial segment of* V*, and let $\pi : M \to \mathrm{V}_\nu$ be elementary. Let $\mathcal{T}$ be a countable*

iteration tree on M. Then there is a maximal branch b through $\mathcal{T}$, and an embedding $\sigma : M_b \to \mathrm{V}_\nu$, so that $\pi = \sigma \circ i^{\mathcal{T}}_b$. (*A branch b whose direct limit can be embedded into V_ν in this way is called* realizable.)

Let M be a model of ZFC. In the (full, length $\omega_1 + 1$) *iteration game* on M players "good" and "bad" collaborate to construct an iteration tree $\mathcal{T}$ of length $\omega_1^{\mathrm{V}} + 1$ on M. "bad" plays all the extenders, and determines the T-predecessor of $\xi + 1$ for each ξ. "good" plays the branches $\{\zeta \mid \zeta \ T \ \lambda\}$ for limit λ, thereby determining the direct limit model M_λ. Note that "good" is also responsible for the final move, which determines $M_{\omega_1^{\mathrm{V}}}$.

If ever a model along the tree is reached which is illfounded then "bad" wins. Otherwise "good" wins. M is (fully) *iterable* if "good" has a winning strategy in this game. An *iteration strategy* for M is a winning strategy for the good player in the iteration game on M.

Notice that if Theorem 2.1 could be strengthened to state that the realizable branch is unique, then repeated applications of the theorem (including a final application over $\mathrm{V}^{\mathrm{col}(\omega,\omega_1)}$ to obtain a branch through a tree of length ω_1^{V}) would demonstrate that countable elementary substructures of rank initial segments of V are iterable. This observation is the key to many of the known iterability proofs, but unfortunately uniqueness fails beyond certain large cardinals. A general proof of iterability would be a great step forward in the study of large cardinals.

A (fine structural) *inner model* is a model of the form $M = \mathrm{L}_\alpha(\vec{E})$, that is the smallest model of set theory containing the ordinals below α and closed under comprehension relative to $\vec{E}$, where $\vec{E}$ is a sequence of extenders, over M or over initial segments of M, satisfying certain coherence requirements. (There are various ways to structure the sequences. For precise definitions see Mitchell–Steel [26] or Zeman [42].) $M = \mathrm{L}_\alpha(\vec{E})$ is an *initial segment* of $N = \mathrm{L}_\beta(\vec{F})$ just in case that $\alpha \leq \beta$ and $\vec{E}$ is an initial segment of $\vec{F}$. Since the extenders in $\vec{E}$ may be extenders not over M but over strict initial segments of M, an iteration tree on M may involve *dropping* to initial segments, that is applying an extender in M_ξ to an initial segment of M_ζ. In such cases the embedding $j_{\zeta,\xi+1}$ acts on an initial segment of M_ζ.

The following fact is the key to the use of iteration trees in the study of inner models:

Fact 2.2 (Comparison). Let M and N be countable inner models. Suppose that M and N are both iterable. Then there are iteration trees $\mathcal{T}$ and $\mathcal{U}$ of countable lengths on M and N respectively, leading to final models M^* and N^*, so that at least one of the following conditions holds:

1. M^* is an initial segment of N^* and there are no drops on the branch of $\mathcal{T}$ leading from M to M^*.

2. N^* is an initial segment of M^* and there are no drops on the branch of $\mathcal{U}$ leading from N to N^*.

The iteration trees $\mathcal{T}$ and $\mathcal{U}$ witnessing Fact 2.2 are constructed inductively. Suppose the construction reached models M_ξ on $\mathcal{T}$ and N_ξ on $\mathcal{U}$. If the extender sequences of M_ξ and N_ξ agree, meaning that they are equal or that one is a strict initial segment of the other, then the construction is over and one of conditions (1) and (2) in Fact 2.2 holds. If the sequences do not agree, let ρ be least so that the extender sequences of M_ξ and N_ξ disagree on the ρth extender. Set E_ξ to be the ρth extender on the sequence of M_ξ, and use this assignment to continue the construction of $\mathcal{T}$, applying E_ξ to M_ζ for the smallest possible ζ, to give rise to $M_{\xi+1}$. Continue $\mathcal{U}$ similarly using the ρth extender on the sequence of N_ξ. These assignments determine the parts of $\mathcal{T}$ and $\mathcal{U}$ corresponding to the bad player's moves in the iteration game. Using the assumption that M and N are iterable, fix iteration strategies Σ and Λ for the two models, and let these strategies determine the remaining elements of $\mathcal{T}$ and $\mathcal{U}$, namely the branches to be used at limit stages.

It is one of the great discoveries of inner model theory that the process described above, of repeatedly forming ultrapowers by disagreeing extenders, terminates, leading therefore to models which are lined-up with their extender sequences in complete agreement. The discovery was first made by Kunen [15] in the context of a single measurable cardinal, where linear iterations suffice. Mitchell [24], [25] developed the framework for models with many measurable cardinals, still using linear iterations. Martin–Steel [22], [23] discovered that in the context of Woodin cardinals the more general (non-linear) iteration trees are both needed and sufficient. Mitchell–Steel [26] used iteration trees, fine structure (see Jensen [8]), and several additional ideas to develop inner models for Woodin cardinals, and reach Fact 2.2 as stated above.

The following folklore claim illustrates a simple application of the comparison process. An inner model M is called a *minimal* model for a sentence θ if M satisfies θ and no strict initial segment of M satisfies θ.

Claim 2.3. *Let M and N be minimal countable inner models for the same sentence θ. Suppose that both M and N are iterable. Then M and N have the same theory.*

Proof sketch. Compare M and N, that is form $\mathcal{T}$ and $\mathcal{U}$ leading to models M^* and N^* which are in complete agreement, using Fact 2.2. Neither one of M^* and N^* can be a strict initial segment of the other, since otherwise the longer of the two will have a strict initial segment satisfying θ. M^* and N^* must therefore be equal. Similar reasoning shows that there can be drops on either side of the comparison. Using the elementarity of the iteration embeddings (from M to M^* along $\mathcal{T}$, and from N to N^* along $\mathcal{U}$) it follows that M has the same theory as M^* and N has the same theory as N^*. Since $M^* = N^*$, M and N have the same theory. □

An inner model M is a *sharp* if its extender sequence has a final element, E^M_{top}, and E^M_{top} is an extender over M. For a sharp M let M^* be the result of iterating E^M_{top} through the countable ordinals, that is set M' equal to the final model of the iteration tree $\mathcal{T}$ defined by the condition $E_\xi = j_{0,\xi}(E^M_{\text{top}})$ and the T-predecessor of $\xi+1$ is ξ for all $\xi < \omega_1$, and let $M^* = M'\|\omega_1$. The set $I = \{j_{0,\xi}(\text{crit}(E^M_{\text{top}})) \mid \xi < \omega_1\}$

is club in ω_1. The ordinals in I are indiscernibles for M^*, in the sense that for any formula φ, and any increasing sequences $\{\alpha_1, \ldots, \alpha_k\}$ and $\{\beta_1, \ldots, \beta_k\}$ in $[I]^k$, $M^* \models \varphi[\alpha_1, \ldots, \alpha_k]$ iff $M^* \models \varphi[\beta_1, \ldots, \beta_k]$. The *theory of k indiscernibles for M*, denoted $\mathrm{Th}_k(M)$, is the set of formulae φ so that $M^* \models \varphi[\alpha_1, \ldots, \alpha_k]$ for some (equivalently all) $\{\alpha_1, \ldots, \alpha_k\} \in [I]^k$.

An argument similar to that of Claim 2.3 shows that if M and N are both minimal iterable sharps for the same sentence θ, then $\mathrm{Th}_k(M) = \mathrm{Th}_k(N)$. The join $\bigoplus_{k<\omega} \mathrm{Th}_k(M)$ is called *the* sharp for θ. The sharp for the sentence "there are n Woodin cardinals" is called the sharp for n Woodin cardinals. The sharp for a tautology is denoted $0^\sharp$. It codes a club of indiscernibles for L.

The existence of $0^\sharp$ follows from the existence of a measurable cardinal. But in general the existence of the sharp for θ does not follow directly from the existence of large cardinals in V. The sharp also requires *iterability*, which is used in an essential way through the appeal to the comparison process in the proof of Claim 2.3. At the level of finitely many Woodin cardinals iterability can be obtained using Theorem 2.1 and additional arguments on the uniqueness of realizable branches, so that the existence of the sharp for n Woodin cardinals follows from the existence in V of n Woodin cardinals and a measurable cardinal above them.

It was noted in Section 1 that proofs of determinacy for pointclasses from $\mathbf{\Pi}^1_1$ onward require large cardinal axioms. To be specific, a proof of determinacy for the pointclass $\mathbf{\Pi}^1_1$ (Theorem 1.3) requires roughly the existence of a measurable cardinal, a proof of determinacy for the pointclass $\mathbf{\Pi}^1_{n+1}$ (Theorem 1.4) requires roughly the existence of n Woodin cardinals and a measurable cardinal above them, and a proof of determinacy for the pointclass of all sets in $\mathrm{L}(\mathbb{R})$ (Theorem 1.5) requires roughly the existence of ω Woodin cardinals and a measurable cardinal above them. But this is only the beginning of the connection between these pointclasses and Woodin cardinals.

Recall that a set A is α-$\mathbf{\Pi}^1_1$ if there is a sequence $\langle A_\xi \mid \xi < \alpha\rangle$ of $\mathbf{\Pi}^1_1$ sets so that $x \in A$ iff the least ξ so that $x \notin A_\xi \vee \xi = \alpha$ is odd. (The hierarchy generated by this definition is the *difference hierarchy* on $\mathbf{\Pi}^1_1$ sets. Note for example that for $\alpha = 2$ the condition states simply that $A = A_0 - A_1$.) The set A is (lightface) α-Π^1_1 if the underlying code for the sequence $\langle A_\xi \mid \xi < \alpha\rangle$ is recursive. A is $<\omega^2$-Π^1_1 if it is α-Π^1_1 for some $\alpha < \omega^2$.

Theorem 2.4 (Martin [21]). *Let B_i $(i < \omega)$ be a recursive enumeration of the $<\omega^2$-Π^1_1 sets. Then each of $0^\sharp$ and $\{i \mid$ player* I *has a winning strategy in $G_\omega(B_i)\}$ is recursive in the other.*

Theorem 2.4 provides a very tight connection between a large cardinal object, $0^\sharp$, and infinite games. For every formula φ there is a $<\omega^2$-Π^1_1 set B so that φ belongs to $0^\sharp$ iff I wins $G_\omega(B)$, and conversely (for every B there is φ).

Theorem 2.5. *Let B_i $(i < \omega)$ be a recursive enumeration of the $\supset^{(n)}(<\omega^2$-$\Pi^1_1)$ sets. Then the sharp for n Woodin cardinals and $\{i \mid$ player* I *has a winning strategy in $G_\omega(B_i)\}$ are each recursive in the other.*

Theorem 2.5 generalizes Theorem 2.4 to $n > 0$. It has two directions. The first states that membership in the sharp for n Woodin cardinals is equivalent to winning a $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ game. This follows from the results of Martin–Steel [23]. Essentially they show that iterability (or more precisely the ability to survive through the comparison process) for minimal sharps for n Woodin cardinals, can be expressed as a $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ statement. The other direction of Theorem 2.5 states that sharps for n Woodin cardinals can discern which player wins a $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ game. This direction follows from the determinacy proof in Neeman [30], [32]. The proof reduces the quantifiers involved in the $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ set to an iteration game on any model which has a sharp for n Woodin cardinals. The reduction takes place inside the model, and the model can tell which player in the $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ game is matched to the good player in the iteration game. Since the sharp is iterable, this player wins the $\Game^{(n)}(<\omega^2\text{-}\Pi^1_1)$ game.

Theorem 2.5 is an indication of the close connections between the study of inner models for Woodin cardinals and the study of projective pointclasses. The connections are tight enough that inner models can be used directly in the study of projective pointclasses, and further up in the study of $\mathrm{L}(\mathbb{R})$ under determinacy.

Theorem 2.6 (Neeman–Woodin, see [30]). *Determinacy for all $\mathbf{\Pi}^1_{n+1}$ sets implies determinacy for all sets in the larger pointclass $\Game^{(n)}(<\omega^2\text{-}\mathbf{\Pi}^1_1)$.*

Theorem 2.7 (Hjorth [6]). *Work in $\mathrm{L}(\mathbb{R})$ assuming AD. Let $\preceq$ be a $\Game(\alpha\text{-}\mathbf{\Pi}^1_1)$ prewellorder with $\alpha < \omega \cdot k$. Then the ordertype of $\preceq$ is smaller than ω_{k+1}.*

Theorem 2.8 (Neeman, Woodin, see [36]). *Assume $\mathsf{AD}^{\mathrm{L}(\mathbb{R})}$. Then it is consistent (with $\mathsf{AD}^{\mathrm{L}(\mathbb{R})}$ and the axiom of choice) that $\boldsymbol{\delta}^1_3 = \omega_2$.*

Theorem 2.6 for $n = 0$ is a combination of the work of Harrington [5], who obtained $0^\sharp$ and its relativized versions for all reals from $\mathbf{\Pi}^1_1$ determinacy, and Martin, who obtained $<\omega^2\text{-}\mathbf{\Pi}^1_1$ determinacy from the sharps. At higher levels Woodin obtained sharps for n Woodin cardinals from $\mathbf{\Pi}^1_{n+1}$ determinacy and Neeman [30] obtained $\Game^{(n)}(<\omega^2\text{-}\mathbf{\Pi}^1_1)$ determinacy from these sharps. Theorem 2.6 had already been proved for odd n by Kechris–Woodin [14], using methods which are purely descriptive set theoretic. For even n the only known proofs involve large cardinals.

Hjorth [6] proved Theorem 2.7 by embedding a given $\Game(<\omega \cdot k\text{-}\mathbf{\Pi}^1_1)$ prewellorder into a directed system of iterates of a sharp for one Woodin cardinal, and proving that the rank of the directed system is smaller than ω_{k+1}. Again, the proof is closely tied up with large cardinals and iteration trees, even though the result is purely descriptive set theoretic.

Theorem 2.8 is proved by collapsing ω_ω to ω_1 over $\mathrm{L}(\mathbb{R})$, so that $(\omega_{\omega+1})^{\mathrm{L}(\mathbb{R})}$, which is equal to $\boldsymbol{\delta}^1_3$ by Theorem 1.11, becomes ω_2 of the generic extension. The forcing to collapse ω_ω involves an ultrafilter on $[\mathcal{P}_{\omega_1}(\omega_\omega)]^{<\omega_1}$, and the construction of this ultrafilter uses a directed system of iterates of fine structural inner models with Woodin cardinals.

3. Larger cardinals, longer games

For $\alpha < \omega_1$ and $B \subset \mathbb{R}^\alpha$ let $G_{\omega\cdot\alpha}(B)$ denote the length $\omega \cdot \alpha$ game with payoff B. Players I and II alternate playing natural numbers in the format of Diagram 2, taking $\omega \cdot \alpha$ moves to produce together a sequence $r = \langle r(\xi) \mid \xi < \omega \cdot \alpha\rangle$ in $\omega^{\omega\cdot\alpha}$. The sequence r may be viewed as an element of $(\omega^\omega)^\alpha = \mathbb{R}^\alpha$. If r belongs to B then player I wins, and otherwise player II wins.

I	$r(0)$		$r(2)$		$\ldots$	$r(2\xi)$		$\ldots\ldots$	
II		$r(1)$		$\ldots$			$r(2\xi+1)$		$\ldots\ldots$

Diagram 2. General format of a transfinite game.

Determinacy for all length ω games with payoff in $\mathcal{D}^{(n)}(<\omega^2\text{-}\Pi^1_1)$ is easily seen to be the same as determinacy for all games of length $\omega \cdot (n+1)$ with payoff in $<\omega^2\text{-}\Pi^1_1$. Theorem 2.5 and the part of Theorem 2.6 dealing with a proof of determinacy from sharps can therefore be rephrased as follows:

Theorem 3.1. *Let B_i $(i < \omega)$ be a recursive enumeration of all the $<\omega^2\text{-}\Pi^1_1$ sets. Suppose that there is an iterable sharp for n Woodin cardinals. Then all length $\omega \cdot (n+1)$ games with payoff in $<\omega^2\text{-}\Pi^1_1$ are determined. Moreover, the sharp for n Woodin cardinals and $\{i \mid$ player* I *wins $G_{\omega\cdot(n+1)}(B_i)\}$ are each recursive in the other.*

The same precise connection between large cardinals and determinacy can be obtained higher up. Theorems 3.2, 3.3, and 3.4 below give several markers along the hierarchies of large cardinals and determinacy, progressively moving upward on both.

Theorem 3.2 (Neeman, Woodin, see [33, Chapter 2]). *Let α be a countable ordinal. Let B_i $(i < \omega)$ be a recursive[2] enumeration of all the $<\omega^2\text{-}\Pi^1_1$ subsets of $\mathbb{R}^{1+\alpha}$. Suppose that there is an iterable sharp for α Woodin cardinals. Then all length $\omega \cdot (1+\alpha)$ games with payoff in $<\omega^2\text{-}\Pi^1_1$ are determined. Moreover, the sharp for α Woodin cardinals and $\{i \mid$ player* I *wins $G_{\omega\cdot(1+\alpha)}(B_i)\}$ are each recursive in the other.*

For $B \subset \omega^{<\omega_1}$ let $G_{\mathrm{adm}}(B)$ be the following game: Players I and II alternate natural numbers as in Diagram 2, continuing until they reach the first ordinal α so that $\mathrm{L}_\alpha[r(\xi) \mid \xi < \alpha]$ is admissible. At that point the game ends. Player I wins if $\langle r(\xi) \mid \xi < \alpha\rangle \in B$, and otherwise player II wins.

The run $\langle r(\xi) \mid \xi < \alpha\rangle$ has the property that for every $\beta < \alpha$, $\mathrm{L}_\beta[r(\xi) \mid \xi < \beta]$ is not admissible. Using this property the run can be coded by a real in a canonical, uniform manner. The payoff set B is said to be Γ *in the codes* just in case that the set of real codes for sequences in B belongs to Γ.

$G_{\mathrm{adm}}(B)$ is a game of *variable countable length*. Its runs are countable, but the length of a particular run depends on the moves made during the run. Each of the

[2] Recursiveness here is relative to a code for α.

players can force the length of the run to be greater than any fixed countable ordinal α, and the determinacy of $G_{\text{adm}}(B)$ for all B in $<\omega^2\text{-}\Pi^1_1$ implies the determinacy of $G_\alpha(B)$ for all B in $<\omega^2\text{-}\Pi^1_1$, for each countable α.

The *Mitchell order* on extenders is the order $E \lhd F$ iff $E \in \text{Ult}(\text{V}, F)$. The Mitchell order of a cardinal κ is the ordertype of the restriction of $\lhd$ to extenders with critical point κ.

Theorem 3.3 (Neeman [34]). *Let B_i $(i < \omega)$ be a recursive enumeration of the subsets of $\omega^{<\omega_1}$ which are $<\omega^2\text{-}\Pi^1_1$ in the codes. Suppose that there is an iterable sharp for the statement "there is a cardinal κ which is a limit of Woodin cardinals and has Mitchell order κ^{++}". Then the games $G_{\text{adm}}(B)$ are determined for all B which are $<\omega^2\text{-}\Pi^1_1$ in the codes. Moreover the sharp and the real $\{i \mid$ player* I *wins $G_{\text{adm}}(B_i)\}$ are each recursive in the other.*

For $B \subset \omega^{<\omega_1}$ let $G_{\text{local}}(\text{L}, B)$ be the following game: Players I and II alternate natural numbers as in Diagram 2, continuing until they reach the first $\alpha > \omega$ so that α is a cardinal in $\text{L}[r(\xi) \mid \xi < \alpha]$. At that point the game ends. Player I wins if $\langle r(\xi) \mid \xi < \alpha\rangle \in B$. Otherwise player II wins. $G_{\text{local}}(\text{L}, B)$ is a *game ending at ω_1 in* L *of the play*. It too is a game of variable countable length.

A code for a run $\langle r(\xi) \mid \xi < \alpha\rangle$ of $G_{\text{local}}(\text{L}, B)$ is simply a pair $\langle w, x\rangle$ where w is a wellorder of ω of ordertype α, $x \in \omega^\omega$, and for each n, $x(n)$ is equal to $r(\xi)$ where ξ is the ordertype of n in w. These codes belong to $\mathcal{P}(\omega \times \omega) \times \omega^\omega$, which is isomorphic to ω^ω. As before, B is said to be Γ *in the codes* just in case that the set of codes for sequences in B belongs to Γ.

Theorem 3.4 (Neeman [33, Chapter 7]). *Let B_i $(i < \omega)$ be a recursive enumeration of the subsets of $\omega^{<\omega_1}$ which are $\partial(<\omega^2\text{-}\Pi^1_1)$ in the codes. Suppose that there is an iterable sharp for the statement "there is a Woodin cardinal which is also a limit of Woodin cardinals". Then the games $G_{\text{local}}(\text{L}, B)$ are determined for all B which are $\partial(<\omega^2\text{-}\Pi^1_1)$ in the codes. Moreover the sharp and the real $\{i \mid$ player* I *wins $G_{\text{local}}(\text{L}, B_i)\}$ are each recursive in the other.*

Remark 3.5. Theorem 3.4 has an interesting corollary, due to Woodin: Suppose that there is an iterable sharp for a Woodin limit of Woodin cardinals. Then it is consistent that all ordinal definable games of length ω_1 are determined. The model witnessing this is of the form $M = \text{L}[x(\xi) \mid \xi < \gamma]$ where $\gamma = \omega_1^M$, and the strategies witnessing determinacy in this model are obtained through uses of Theorem 3.4 on games ending at ω_1 in L of the play. For a complete proof of the corollary see Neeman [33, 7F.13–15].

Remark 3.6. There is another interesting game that comes up in the proof of Theorem 3.4. For a partial function $f : \mathbb{R} \to \omega$ and a set $B \subset \omega^{<\omega_1}$ let $G_{\text{cont}}(f, B)$ be the following game: Players I and II alternate natural numbers as in Diagram 2. In addition, after each block of ω moves they write a natural numbers on a "side board". Let $x_\alpha = \langle r(\omega \cdot \alpha + i) \mid i < \omega\rangle$ be the αth block of moves. The natural number

they write following this block is $n_\alpha = f(x_\alpha)$. They continue playing until reaching the first α so that $x_\alpha \notin \mathrm{dom}(f)$ or $n_\alpha \in \{n_\beta \mid \beta < \alpha\}$ (meaning that the natural number written after block α is a repetition of a number written previously). At that point the game ends, player I wins if $\langle r(\xi) \mid \xi < \omega \cdot \alpha + \omega\rangle \in B$, and player II wins otherwise. The large cardinal strength of determinacy for these games is roughly a cardinal κ which is $\delta + 1$-strong for some Woodin cardinal $\delta > \kappa$ (see Neeman [33, Chapter 3]), and the determinacy proof for these games is a precursor to the use of extenders overlapping Woodin cardinals in the proof of Theorem 3.4.

Determinacy in Theorems 3.2, 3.3, and 3.4 is proved by reducing the long game to an iteration game on the given model. The reduction, which uses the large cardinals of the model, matches one of the players in the long game to the good player in the iteration game. In effect it converts the iteration strategy for the model into a winning strategy for this player in the long game. Determinacy therefore rests on the existence of *iterable* models; the existence of large cardinals by itself is not directly sufficient.

In the case of Theorems 3.2 and 3.3, the long game is reduced to an iteration game of a specific format, involving only linear compositions of iteration trees of length ω. The fact that the good player can win games of this format, on countable models which embed into rank initial segments of V, follows directly from Theorem 2.1. The determinacy proved in Theorems 3.2 and 3.3 therefore follows from just the assumptions of large cardinals in V: α Woodin cardinals and a measurable cardinal above them in the case of Theorem 3.2, and a measurable cardinal above a cardinal κ so that $o(\kappa) = \kappa^{++}$ and κ is a limit of Woodin cardinals in the case of Theorem 3.3.

The iteration game generated by the proof of Theorem 3.4 is as complicated as the full iteration game, and Theorem 2.1 by itself is not enough to produce a winning strategy for the good player in this game. Still, by Neeman [31], the existence of an iterable model satisfying the large cardinal assumptions of Theorem 3.4 follows from the existence of the large cardinals, a Woodin limit of Woodin cardinals and a measurable cardinal above it, in V.

Theorems 3.2, 3.3, and 3.4 extend the precise connection between determinacy and inner models to levels of games of variable countable lengths, and Woodin limits of Woodin cardinals. It is generally believed that the large cardinal hierarchy is rich enough to calibrate the strength of all natural statements. Could determinacy provide a rich enough hierarchy to match the full extent of the large cardinal hierarchy? If not, how far does determinacy reach? How far does the hierarchy of long games reach? We are very far from answers to these questions.

Let θ be a large cardinal assumption at or below the existence of a superstrong cardinal. (Beyond the level of superstrong cardinals there are problems with the comparison process and Claim 2.3.) The comparison process provides the best clues in the search for long games strong enough to match θ in the sense of Theorems 3.2, 3.3, and 3.4: If a particular format of long games subsumes the iteration games appearing in the comparison of minimal models of θ, then the associated game quantifier is strong enough to define the sharp for θ.

The following format is strong enough to subsume the full iteration game, and therefore all iteration games appearing in all comparisons of all inner models up to superstrong cardinals. Let $\mathcal{L}^+$ denote the language of set theory with an added unary relation symbol $\dot{r}$, and let $\varphi(\alpha, \beta)$ be a formula in $\mathcal{L}^+$. Define $G_{\text{club},2}(\varphi)$ to be the following game: Players I and II alternate playing ω_1 natural numbers in the manner of Diagram 2, producing together a sequence $\langle r(\xi) \mid \xi < \omega_1 \rangle$ in ω^{ω_1}. If there is a club $C \subset \omega_1$ so that $\langle \mathrm{L}_{\omega_1}[r], r \rangle \models \varphi[\alpha, \beta]$ for all $\alpha < \beta$ both in C then player I wins, and otherwise player II wins. (A quick word on notation: r formally is a set of pairs in $\omega_1 \times \omega$. $\langle \mathrm{L}_{\omega_1}[r], r \rangle \models \varphi$ iff φ holds in $\mathrm{L}_{\omega_1}[r]$ with appearances of the predicate $\dot{r}$ in φ interpreted by r.)

The number 2 in $G_{\text{club},2}(\varphi)$ refers to the number of free variables in φ. Games $G_{\text{club},k}(\varphi)$, for $k \neq 2$ in ω and φ a formula with k free variables, can defined similarly. All the definitions can be relativized to a real x by replacing $\mathrm{L}_{\omega_1}[r]$ with $\mathrm{L}_{\omega_1}[r, x]$ and letting φ take x as a parameter. They can be relativized to a set of reals A by replacing L with L^A and allowing φ to take an additional predicate interpreted by $A \cap \mathrm{L}^A_{\omega_1}[r]$.

The full iteration game on a countable model M can be recast as a game of the form $G_{\text{club},2}(\varphi)$ relativized to a real coding M. Woodin [41] connects the determinacy of the games $G_{\text{club},k}(\varphi)$ and their relativizations to Σ^2_2 absoluteness under the combinatorial principle generic diamond ($\diamondsuit_G$). Determinacy for the games $G_{\text{club},2}(\varphi)$ is not provable from large cardinals, by Larson [17], but it may be provable from large cardinals and $\diamondsuit_G$. Unfortunately the games are too strong to be handled by current methods in proofs of determinacy, precisely because they are strong enough to subsume the full iteration game. If there were a match for $G_{\text{club},2}$ similar to the matches in Theorems 2.4, 2.5, 3.2, 3.3, and 3.4, then the large cardinal involved would have to be stronger than a superstrong, far beyond the level of Woodin cardinals.

The following format produces a weaker game. Let $k < \omega$. Let $\vec{S} = \langle S_a \mid a \in [\omega_1]^{<k} \rangle$ be a collection of mutually disjoint stationary subsets of ω_1, with a stationary set S_a associated to each tuple $a \in [\omega_1]^{<k}$. Let $[\vec{S}]$ denote the set $\{\langle \alpha_0, \dots, \alpha_{k-1} \rangle \in [\omega_1]^k \mid (\forall i < k)\ \alpha_i \in S_{\langle \alpha_0, \dots, \alpha_{i-1} \rangle}\}$. Let $\varphi(x_0, \dots, x_{k-1})$ be a formula in $\mathcal{L}^+$. Define $G_{\omega_1,k}(\vec{S}, \varphi)$ to be the following game: Players I and II alternate playing ω_1 natural numbers in the manner of Diagram 2, producing together a sequence $r \in \omega^{\omega_1}$. If there is a club $C \subset \omega_1$ so that $\langle \mathrm{L}_{\omega_1}[r], r \rangle \models \varphi[\alpha_0, \dots, \alpha_{k-1}]$ for all $\langle \alpha_0, \dots, \alpha_{k-1} \rangle \in [\vec{S}] \cap [C]^k$ then player I wins the run r. If there is a club $C \subset \omega_1$ so that $\langle \mathrm{L}_{\omega_1}[r], r \rangle \models \neg\varphi[\alpha_0, \dots, \alpha_{k-1}]$ for all $\langle \alpha_0, \dots, \alpha_{k-1} \rangle \in [\vec{S}] \cap [C]^k$ then player II wins r. If neither condition holds then both players lose.

Note that the two winning conditions in the definition of $G_{\omega_1,k}(\vec{S}, \varphi)$ cannot both hold. This uses the fact that each of the sets S_a is stationary in ω_1, and the demand in the conditions that C must be club in ω_1. Thus at most one player wins each run of $G_{\omega_1,k}(\vec{S}, \varphi)$. For $k > 0$ it may well be that neither one of the winning conditions holds. So there may well be runs of $G_{\omega_1,k}(\vec{S}, \varphi)$ which are won by neither player. Determinacy for $G_{\omega_1,k}(\vec{S}, \varphi)$ is defined in the stronger of the two possible senses. The game is determined if one of the players has a winning strategy; a strategy which

merely avoids losing is not enough.

Recall that a sharp for θ is an inner model M with a final extender E^M_{top}, so that E^M_{top} is an extender over M and $M \models \theta$. Let θ be the sentence "crit(E^M_{top}) is a Woodin cardinal". The minimal iterable sharp for θ, if it exists, is denoted 0^W. Recall that iterating out the top extender of a sharp M produces a model M^* and a club $I \subset \omega_1$ of indiscernibles for M^*, consisting of the images of crit(E^M_{top}) under the iteration embeddings. In the case of $M = 0^W$, the ordinals in I are Woodin cardinals of M^*. The existence of 0^W thus implies the existence of an iterable model with a club of indiscernible Woodin cardinals, and in fact the two are equivalent.

Remark 3.7. Iterability for countable elementary substructures of V is not known at the level of 0^W – the strongest results in this direction are the ones of Neeman [31], reaching to the level of Woodin limits of Woodin cardinals – and the existence of 0^W is not known to follow from large cardinals in V.

Theorem 3.8 (Neeman [35]). *Suppose that 0^W exists. Then the games $G_{\omega_1,k}(\vec{S}, \varphi)$ are determined, for all $\vec{S}$, k, and φ.*

There are two parameters determining the payoff of the game $G_{\omega_1,k}(\vec{S}, \varphi)$. One is the formula φ and the number k of its free variables. The other is the sequence $\vec{S}$. The formula φ, or the formula φ and the real x in the case of games relativized to a real, is the definable part of the payoff condition, analogous to the $<\omega^2$-Π^1_1 set, or more precisely to its recursive definition, in Theorems 3.2, 3.3, and 3.4. The sequence $\vec{S}$ consists of disjoint stationary sets, and this makes it highly non-definable. It has no parallel in Theorems 3.2, 3.3, and 3.4. It is necessary in Theorem 3.8, and the winning strategy in $G_{\omega_1,k}(\vec{S}, \varphi)$ depends on $\vec{S}$. But which of the players has a winning strategy is determined independently of $\vec{S}$:

Theorem 3.9 (Neeman [35]). *Suppose that 0^W exists. Let $\vec{S} = \langle S_a \mid a \in [\omega_1]^{<k} \rangle$ and $\vec{S}^* = \langle S^*_a \mid a \in [\omega_1]^{<k} \rangle$ be two sequences of mutually disjoint stationary subsets of ω_1. Then player* I (*respectively* II) *has a winning strategy in $G_{\omega_1,k}(\vec{S}, \varphi)$ iff she has a winning strategy in $G_{\omega_1,k}(\vec{S}^*, \varphi)$.*

Define $\beth_{\omega_1}(k, \varphi)$ to be 1 if player I has a winning strategy in $G_{\omega_1,k}(\vec{S}, \varphi)$ for some, and using Theorem 3.9 equivalently for all, $\vec{S}$. Define $\beth_{\omega_1}(k, \varphi)$ to be equal to 0 otherwise.

Theorem 3.10 (Neeman [35]). *Suppose that 0^W exists. Then $\{\langle k, \varphi \rangle \mid \beth_{\omega_1}(k, \varphi) = 1\}$ and 0^W are each recursive in the other.*

Theorems 3.8 and 3.10 establish the same precise connection between 0^W and games of length ω_1 that exists between $0^\sharp$ and $<\omega^2$-Π^1_1 games of length ω. They provide another step, the first to reach games of length ω_1, in the project of matching the hierarchy of large cardinals with the hierarchy of long games.

References

[1] Addison, John W., and Moschovakis, Yiannis N., Some consequences of the axiom of definable determinateness. *Proc. Nat. Acad. Sci. U.S.A.* **59** (1968), 708–712.

[2] Blackwell, David, Infinite games and analytic sets. *Proc. Nat. Acad. Sci. U.S.A.* **58** (1967), 1836–1837.

[3] Davis, Morton, Infinite games of perfect information. In *Advances in game theory*, Princeton University Press, Princeton, N.J., 1964, 85–101.

[4] Gale, David, and Stewart, Frank M., Infinite games with perfect information. In *Contributions to the theory of games*, vol. 2, Ann. of Math. Stud. 28, Princeton University Press, Princeton, N. J., 1953, 245–266.

[5] Harrington, Leo, Analytic determinacy and $0^\sharp$. *J. Symbolic Logic* **43** (4) (1978), 685–693.

[6] Hjorth, Greg, A boundedness lemma for iterations. *J. Symbolic Logic* **66** (3) (2001), 1058–1072.

[7] Jackson, Steve, Structural consequences of AD. In *Handbook of Set Theory*, to appear.

[8] Jensen, R. Björn, The fine structure of the constructible hierarchy. With a section by Jack Silver. *Ann. Math. Logic* **4** (1972), 229–308; erratum, ibid. **4** (1972), 443.

[9] Kechris, Alexander S., AD and projective ordinals. In *Cabal Seminar 76–77*, Lecture Notes in Math. 689, Springer-Verlag, Berlin 1978, 91–132.

[10] Kechris, Alexander S., and Martin, Donald A. (eds.), *Cabal Seminar 77–79*. Lecture Notes in Math. 839, Springer-Verlag, Berlin 1981.

[11] Kechris, Alexander S., Martin, Donald A., and Moschovakis, Yiannis N. (eds.), *Cabal seminar 79–81*. Lecture Notes in Math. 1019, Springer-Verlag, Berlin 1983.

[12] Kechris, Alexander S., Martin, Donald A., and Steel, John R. (eds.), *Cabal Seminar 81–85*. Lecture Notes in Math. 1333, Springer-Verlag, Berlin 1988.

[13] Kechris, Alexander S., and Moschovakis, Yiannis N. (eds.), *Cabal Seminar 76–77*. Lecture Notes in Math. 689, Springer-Verlag, Berlin 1978.

[14] Kechris, Alexander S., and Woodin, W. Hugh, Equivalence of partition properties and determinacy. *Proc. Nat. Acad. Sci. U.S.A.* **80** (6i.) (1983), 1783–1786.

[15] Kunen, Kenneth, Some applications of iterated ultrapowers in set theory. *Ann. Math. Logic* **1** (1970), 179–227.

[16] Kuratowski, Casimir, Sur les théorèmes de séparation dans la théorie des ensembles. *Fund. Math.* **26** (1936), 183–191.

[17] Larson, Paul B., The canonical function game. *Arch. Math. Logic* **44** (2005), 817–827.

[18] Martin, Donald A., The axiom of determinateness and reduction principles in the analytical hierarchy. *Bull. Amer. Math. Soc.* **74** (1968), 687–689.

[19] Martin, Donald A., Measurable cardinals and analytic games. *Fund. Math.* **66** (1969/1970), 287–291 .

[20] Martin, Donald A., Borel determinacy. *Ann. of Math.* (2) **102** (2) (1975), 363–371.

[21] Martin, Donald A., The largest countable this, that, and the other. In *Cabal seminar 79–81*, Lecture Notes in Math. 1019, Springer-Verlag, Berlin 1983, 97–106.

[22] Martin, Donald A., and Steel, John R., A proof of projective determinacy. *J. Amer. Math. Soc.* **2** (1) (1989), 71–125.

[23] Martin, Donald A., and Steel, John R., Iteration trees. *J. Amer. Math. Soc.* **7** (1) (1994), 1–73.

[24] Mitchell, William J., Sets constructible from sequences of ultrafilters. *J. Symbolic Logic* **39** (1974), 57–66.

[25] Mitchell, William J., Sets constructed from sequences of measures: revisited. *J. Symbolic Logic* **48** (3) (1983), 600–609.

[26] Mitchell, William J., and Steel, John R., *Fine structure and iteration trees*. Lecture Notes in Logic 3, Springer-Verlag, Berlin 1994.

[27] Moschovakis, Yiannis N., *Descriptive set theory*. Stud. Logic Found. Math. 100, North-Holland Publishing Co., Amsterdam 1980.

[28] Mycielski, Jan, and Steinhaus, Hugo, A mathematical axiom contradicting the axiom of choice. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **10** (1962), 1–3.

[29] Mycielski, Jan, and Świerczkowski, Stanisław, On the Lebesgue measurability and the axiom of determinateness. *Fund. Math.* **54** (1964), 67–71.

[30] Neeman, Itay, Optimal proofs of determinacy. *Bull. Symbolic Logic* **1** (3) (1995), 327–339.

[31] Neeman, Itay, Inner models in the region of a Woodin limit of Woodin cardinals. *Ann. Pure Appl. Logic* **116** (1-3) (2002), 67–155.

[32] Neeman, Itay, Optimal proofs of determinacy. II. *J. Math. Log.* **2** (2) (2002), 227–258.

[33] Neeman, Itay, *The determinacy of long games*. De Gruyter Ser. Log. Appl. 7, Walter de Gruyter, Berlin 2004.

[34] Neeman, Itay, Determinacy for games ending at the first admissible relative to the play. *J. Symbolic Logic* **71** (2006), 425–459.

[35] Neeman, Itay, Games of length ω_1. To appear.

[36] Neeman, Itay, Inner models and ultrafilters in L($\mathbb{R}$). To appear.

[37] Oxtoby, John C., The Banach-Mazur game and Banach category theorem. In *Contributions to the theory of games*, vol. 3, Ann. of Math. Stud. 39, Princeton University Press, Princeton, N. J., 1957, 159–163.

[38] Steel, John R., and Van Wesep, Robert, Two consequences of determinacy consistent with choice. *Trans. Amer. Math. Soc.* **272** (1) (1982), 67–85.

[39] Woodin, W. Hugh, Some consistency results in ZFC using AD. In *Cabal seminar 79–81*, Lecture Notes in Math. 1019, Springer-Verlag, Berlin 1983, 172–198.

[40] Woodin, W. Hugh, Supercompact cardinals, sets of reals, and weakly homogeneous trees. *Proc. Nat. Acad. Sci. U.S.A.* **85** (18) (1988), 6587–6591.

[41] Woodin, W. Hugh, Beyond Σ^2_1 absoluteness. *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. I, Higher Ed. Press, Beijing 2002, 515–524.

[42] Zeman, Martin, *Inner models and large cardinals*. De Gruyter Ser. Log. Appl. 5, Walter de Gruyter, Berlin 2002.

Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095-1555, U.S.A.

E-mail: ineeman@math.ucla.edu

The art of ordinal analysis

Michael Rathjen

Abstract. Ordinal analysis of theories is a core area of proof theory whose origins can be traced back to Hilbert's programme – the aim of which was to lay to rest all worries about the foundations of mathematics once and for all by securing mathematics via an absolute proof of consistency. Ordinal-theoretic proof theory came into existence in 1936, springing forth from Gentzen's head in the course of his consistency proof of arithmetic. The central theme of ordinal analysis is the classification of theories by means of transfinite ordinals that measure their 'consistency strength' and 'computational power'. The so-called *proof-theoretic ordinal* of a theory also serves to characterize its provably recursive functions and can yield both conservation and combinatorial independence results.

This paper intends to survey the development of "ordinally informative" proof theory from the work of Gentzen up to more recent advances in determining the proof-theoretic ordinals of strong subsystems of second order arithmetic.

Mathematics Subject Classification (2000). Primary 03F15, 03F05, 03F35; Secondary 03F03, 03-03.

Keywords. Proof theory, ordinal analysis, ordinal representation systems, proof-theoretic strength.

1. Introduction

Ordinal analysis of theories is a core area of proof theory. The origins of proof theory can be traced back to the second problem on Hilbert's famous list of problems (presented at the Second International Congress in Paris on August 8, 1900), which called for a proof of consistency of the arithmetical axioms of the reals. Hilbert's work on axiomatic geometry marked the beginning of his live-long interest in the axiomatic method. For geometry, he solved the problem of consistency by furnishing arithmetical-analytical interpretations of the axioms, thereby reducing the question of consistency to the consistency of the axioms for real numbers. The consistency of the latter system of axioms is therefore the ultimate problem for the foundations of mathematics.

Which axioms for real numbers Hilbert had in mind in his problem was made precise only when he took up logic full scale in the 1920s and proposed a research programme with the aim of providing mathematics with a secure foundation. This was to be accomplished by first formalizing logic and mathematics in their entirety, and then showing that these formalizations are consistent, that is to say free of contra-

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

dictions. Strong restrictions were placed on the methods to be applied in consistency proofs of axiom systems for mathematics: namely, these methods were to be completely *finitistic* in character. The proposal to obtain finitistic consistency proofs of axiom systems for mathematics came to be called *Hilbert's Programme*.

Hilbert's Programme is a reductive enterprise with the aim of showing that whenever a 'real' proposition can be proved by 'ideal' means, it can also be proved by 'real', finitistic means. However, Hilbert's so-called formalism was not intended to eliminate nonconstructive existence proofs in the practice of mathematics, but to vindicate them.

In the 1920s, Ackermann and von Neumann, in pursuit of Hilbert's Programme, were working on consistency proofs for arithmetical systems. Ackermann's 1924 dissertation gives a consistency proof for a second-order version of primitive recursive arithmetic which explicitly uses a finitistic version of transfinite induction up to the ordinal ω^{ω^ω}. The employment of transfinite induction on ordinals in consistency proofs came explicitly to the fore in Gentzen's 1936 consistency proof for Peano arithmetic, **PA**. This proof led to the assignment of a *proof-theoretic ordinal* to a theory. This so-called *ordinal analysis* of theories allows one to classify theories by means of transfinite ordinals that measure their 'consistency strength' and 'computational power'.

The subject of this paper is the development of ordinal analysis from the work of Gentzen up to very recent advances in determining the proof-theoretic ordinals of strong subsystems of second order arithmetic.

1.1. Gentzen's result. The most important structure in mathematics is arguably the structure of the natural numbers $\mathfrak{N} = \big(\mathbb{N}; 0^{\mathfrak{N}}, 1^{\mathfrak{N}}, +^{\mathfrak{N}}, \times^{\mathfrak{N}}, E^{\mathfrak{N}}, <^{\mathfrak{N}}\big)$, where $0^{\mathfrak{N}}$ denotes zero, $1^{\mathfrak{N}}$ denotes the number one, $+^{\mathfrak{N}}$, $\times^{\mathfrak{N}}$, $E^{\mathfrak{N}}$ denote the successor, addition, multiplication, and exponentiation function, respectively, and $<^{\mathfrak{N}}$ stands for the less-than relation on the natural numbers. In particular, $E^{\mathfrak{N}}(n, m) = n^m$.

Many of the famous theorems and problems of mathematics such as Fermat's and Goldbach's conjecture, the Twin Prime conjecture, and Riemann's hypothesis can be formalized as sentences of the language of $\mathfrak{N}$ and thus concern questions about the structure $\mathfrak{N}$.

Definition 1.1. A theory designed with the intent of axiomatizing the structure $\mathfrak{N}$ is *Peano arithmetic*, **PA**. The language of **PA** has the predicate symbols $=$, $<$, the function symbols $+$, $\times$, E (for addition, multiplication, exponentiation) and the constant symbols 0 and 1. The *Axioms of* **PA** comprise the usual equations and laws for addition, multiplication, exponentiation, and the less-than relation. In addition, **PA** has the *Induction Scheme*

$$(\mathbf{IND}) \qquad \varphi(0) \wedge \forall x[\varphi(x) \to \varphi(x+1)] \to \forall x \varphi(x)$$

for all formulae φ of the language of **PA**.

Gentzen showed that transfinite induction up to the ordinal

$$\varepsilon_0 = \sup\{\omega, \omega^\omega, \omega^{\omega^\omega}, \dots\} = \text{least } \alpha.\ \omega^\alpha = \alpha$$

suffices to prove the consistency of **PA**. To appreciate Gentzen's result it is pivotal to note that he applied transfinite induction up to ε_0 solely to elementary computable predicates and besides that his proof used only finitistically justified means. Hence, a more precise rendering of Gentzen's result is

$$\mathbf{F} + \text{EC-TI}(\varepsilon_0) \vdash \mathsf{Con}(\mathbf{PA}); \tag{1}$$

here **F** signifies a theory that embodies only finitistically acceptable means, EC-TI(ε_0) stands for transfinite induction up to ε_0 for elementary computable predicates, and $\mathsf{Con}(\mathbf{PA})$ expresses the consistency of **PA**. Gentzen also showed that his result was the best possible in that **PA** proves transfinite induction up to α for arithmetic predicates for any $\alpha < \varepsilon_0$. The compelling picture conjured up by the above is that the non-finitist part of **PA** is encapsulated in EC-TI(ε_0) and therefore "measured" by ε_0, thereby tempting one to adopt the following definition of *proof-theoretic ordinal* of a theory T:

$$|T|_{\mathsf{Con}} = \text{least } \alpha.\ \mathbf{F} + \text{EC-TI}(\alpha) \vdash \mathsf{Con}(T). \tag{2}$$

In the above, many notions were left unexplained. We will now consider them one by one. The *elementary computable functions* are exactly the Kalmar *elementary functions,* i.e. the class of functions which contains the successor, projection, zero, addition, multiplication, and modified subtraction functions and is closed under composition and bounded sums and products. A predicate is elementary computable if its characteristic function is elementary computable.

According to an influential analysis of finitism due to W.W. Tait, finististic reasoning coincides with a system known as *primitive recursive arithmetic*. For the purposes of ordinal analysis, however, it suffices to identify **F** with an even more restricted theory known as *Elementary Recursive Arithmetic*, **ERA**. **ERA** is a weak subsystem of **PA** having the same defining axioms for $+, \times, E, <$ but with induction restricted to elementary computable predicates.

In order to formalize EC-TI(α) in the language of arithmetic we should first discuss ordinals and the representation of particular ordinals α as relations on $\mathbb{N}$.

Definition 1.2. A set A equipped with a total ordering $\prec$ (i.e. $\prec$ is transitive, irreflexive, and $\forall x, y \in A\,[x \prec y \vee x = y \vee y \prec x]$) is a *wellordering* if every non-empty subset X of A contains a $\prec$-least element, i.e. $(\exists u \in X)(\forall y \in X)[u \prec y \vee u = y]$.

An *ordinal* is a transitive set wellordered by the elementhood relation $\in$.

Fact 1.3. Every wellordering $(A, \prec)$ is order isomorphic to an ordinal $(\alpha, \in)$.

Ordinals are traditionally denoted by lower case Greek letters $\alpha, \beta, \gamma, \delta, \dots$ and the relation $\in$ on ordinals is notated simply by $<$. The operations of addition, multiplication, and exponentiation can be defined on all ordinals, however, addition and multiplication are in general not commutative.

We are interested in representing specific ordinals α as relations on $\mathbb{N}$. In essence Cantor [10] defined the first ordinal representation system in 1897. Natural ordinal representation systems are frequently derived from structures of the form

$$\mathfrak{A} = \langle \alpha, f_1, \ldots, f_n, <_\alpha \rangle \tag{3}$$

where α is an ordinal, $<_\alpha$ is the ordering of ordinals restricted to elements of α and the f_i are functions

$$f_i : \underbrace{\alpha \times \cdots \times \alpha}_{k_i \text{ times}} \longrightarrow \alpha$$

for some natural number k_i.

$$\mathbb{A} = \langle A, g_1, \ldots, g_n, \prec \rangle \tag{4}$$

is a *computable* (or *recursive*) *representation* of $\mathfrak{A}$ if the following conditions hold:

1. $A \subseteq \mathbb{N}$ and A is a computable set.
2. $\prec$ is a computable total ordering on A and the functions g_i are computable.
3. $\mathfrak{A} \cong \mathbb{A}$, i.e. the two structures are isomorphic.

Theorem 1.4 (Cantor, 1897). *For every ordinal $\beta > 0$ there exist unique ordinals $\beta_0 \geq \beta_1 \geq \cdots \geq \beta_n$ such that*

$$\beta = \omega^{\beta_0} + \cdots + \omega^{\beta_n}. \tag{5}$$

The representation of β in (5) is called the *Cantor normal form*. We shall write $\beta =_{CNF} \omega^{\beta_1} + \cdots + \omega^{\beta_n}$ to convey that $\beta_0 \geq \beta_1 \geq \cdots \geq \beta_k$.

ε_0 denotes the least ordinal $\alpha > 0$ such that $(\forall \beta < \alpha)\, \omega^\beta < \alpha$. ε_0 can also be described as the least ordinal α such that $\omega^\alpha = \alpha$.

Ordinals $\beta < \varepsilon_0$ have a Cantor normal form with exponents $\beta_i < \beta$ and these exponents have Cantor normal forms with yet again smaller exponents. As this process must terminate, ordinals $< \varepsilon_0$ can be coded by natural numbers. For instance a coding function

$$\ulcorner \, . \, \urcorner : \varepsilon_0 \longrightarrow \mathbb{N}$$

could be defined as follows:

$$\ulcorner \alpha \urcorner = \begin{cases} 0 & \text{if } \alpha = 0, \\ \langle \ulcorner \alpha_1 \urcorner, \ldots, \ulcorner \alpha_n \urcorner \rangle & \text{if } \alpha =_{CNF} \omega^{\alpha_1} + \cdots + \omega^{\alpha_n} \end{cases}$$

where $\langle k_1, \ldots, k_n \rangle := 2^{k_1+1} \ldots p_n^{k_n+1}$ with p_i being the ith prime number (or any other coding of tuples). Further define:

$$A_0 := \text{range of } \ulcorner . \urcorner, \quad \ulcorner \alpha \urcorner \prec \ulcorner \beta \urcorner :\Leftrightarrow \alpha < \beta$$
$$\ulcorner \alpha \urcorner \hat{+} \ulcorner \beta \urcorner := \ulcorner \alpha + \beta \urcorner, \quad \ulcorner \alpha \urcorner \hat{:} \ulcorner \beta \urcorner := \ulcorner \alpha \cdot \beta \urcorner, \quad \hat{\omega}^{\ulcorner \alpha \urcorner} := \ulcorner \omega^\alpha \urcorner.$$

Then

$$\langle \varepsilon_0, +, \cdot, \delta \mapsto \omega^\delta, < \rangle \cong \langle A_0, \hat{+}, \hat{\cdot}, x \mapsto \hat{\omega}^x, \prec \rangle.$$

$A_0, \hat{+}, \hat{\cdot}, x \mapsto \hat{\omega}^x, \prec$ are computable (recursive), in point of fact, they are all elementary computable.

Finally, we can spell out the scheme EC-TI(ε_0) in the language of **PA**:

$$\forall x\ [\forall y\ (y \prec x \to P(y)) \to P(x)] \to \forall x\, P(x)$$

for all elementary computable predicates P.

1.2. Cut Elimination: Gentzen's Hauptsatz. In the consistency proof, Gentzen used his sequent calculus and employed the technique of *cut elimination*. As this is a tool of utmost importance in proof theory and ordinal analysis, a rough outline of the underlying ideas will be discussed next.

The most common logical calculi are *Hilbert-style* systems. They are specified by delineating a collection of schematic logical axioms and some inference rules. The choice of axioms and rules is more or less arbitrary, only subject to the desire to obtain a *complete* system (in the sense of Gödel's completeness theorem). In model theory it is usually enough to know that there is a complete calculus for first order logic as this already entails the compactness theorem.

There are, however, proof calculi without this arbitrariness of axioms and rules. The *natural deduction calculus* and the *sequent calculus* were both invented by *Gentzen*. Both calculi are pretty illustrations of the symmetries of logic. The sequent calculus since is a central tool in ordinal analysis and allows for generalizations to so-called infinitary logics. Gentzen's main theorem about the sequent calculus is the *Hauptsatz*, i.e. *the cut elimination theorem*.

A *sequent* is an expression $\Gamma \Rightarrow \Delta$ where Γ and Δ are finite sequences of formulae $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$, respectively. We also allow for the possibility that Γ or Δ (or both) are empty. The empty sequence will be denoted by $\emptyset$. $\Gamma \Rightarrow \Delta$ is read, informally, as Γ yields Δ or, rather, the *conjunction* of the A_i yields the *disjunction* of the B_j. In particular, we have:

- If Γ is empty, the sequent asserts the disjunction of the B_j.
- If Δ is empty, it asserts the negation of the conjunction of the A_i.
- if Γ and Δ are both empty, it asserts the *impossible*, i.e. a *contradiction*.

We use upper case Greek letters $\Gamma, \Delta, \Lambda, \Theta, \Xi \ldots$ to range over finite sequences of formulae. $\Gamma \subseteq \Delta$ means that every formula of Γ is also a formula of Δ.

Next we list the axioms and rules of the sequent calculus.

- *Identity Axiom:*

  $$A \Rightarrow A$$

  where A is any formula. In point of fact, one could limit this axiom to the case of atomic formulae A.

- *Cut Rule:*

$$\frac{\Gamma \Rightarrow \Delta, A \qquad A, \Lambda \Rightarrow \Theta}{\Gamma, \Lambda \Rightarrow \Delta, \Theta}\ \mathrm{Cut}$$

The formula A is called the *cut formula* of the inference.

- *Structural Rules:*

$$\frac{\Gamma \Rightarrow \Delta}{\Gamma' \Rightarrow \Delta'} \qquad \text{if } \Gamma \subseteq \Gamma',\ \Delta \subseteq \Delta'.$$

A special case of the structural rule, known as *contraction*, occurs when the lower sequent has fewer occurrences of a formula than the upper sequent. For instance, $A, \Gamma \Rightarrow \Delta, B$ follows structurally from $A, A, \Gamma \Rightarrow \Delta, B, B$.

- *Rules for Logical Operations:*

Left	Right
$\dfrac{\Gamma \Rightarrow \Delta, A}{\neg A, \Gamma \Rightarrow \Delta}$	$\dfrac{B, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \neg B}$
$\dfrac{\Gamma \Rightarrow \Delta, A \qquad B, \Lambda \Rightarrow \Theta}{A \to B, \Gamma, \Lambda \Rightarrow \Delta, \Theta}$	$\dfrac{A, \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \to B}$
$\dfrac{A, \Gamma \Rightarrow \Delta}{A \wedge B, \Gamma \Rightarrow \Delta} \quad \dfrac{B, \Gamma \Rightarrow \Delta}{A \wedge B, \Gamma \Rightarrow \Delta}$	$\dfrac{\Gamma \Rightarrow \Delta, A \qquad \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \wedge B}$
$\dfrac{A, \Gamma \Rightarrow \Delta \qquad B, \Gamma \Rightarrow \Delta}{A \vee B, \Gamma \Rightarrow \Delta}$	$\dfrac{\Gamma \Rightarrow \Delta, A}{\Gamma \Rightarrow \Delta, A \vee B} \quad \dfrac{\Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \vee B}$
$\dfrac{F(t), \Gamma \Rightarrow \Delta}{\forall x\, F(x), \Gamma \Rightarrow \Delta}\ \forall \mathrm{L}$	$\dfrac{\Gamma \Rightarrow \Delta, F(a)}{\Gamma \Rightarrow \Delta, \forall x\, F(x)}\ \forall \mathrm{R}$
$\dfrac{F(a), \Gamma \Rightarrow \Delta}{\exists x\, F(x), \Gamma \Rightarrow \Delta}\ \exists \mathrm{L}$	$\dfrac{\Gamma \Rightarrow \Delta, F(t)}{\Gamma \Rightarrow \Delta, \exists x\, F(x)}\ \exists \mathrm{R}$

In $\forall$L and $\exists$R, t is an arbitrary term. The variable a in $\forall$R and $\exists$L is an *eigenvariable* of the respective inference, i.e. a is not to occur in the *lower sequent*.

In the rules for logical operations, the formulae highlighted in the premisses are called the *minor formulae* of that inference, while the formula highlighted in the conclusion is the *principal formula* of that inference. The other formulae of an inference are called *side formulae*.

A *proof* (also known as *deduction* or *derivation*) $\mathcal{D}$ is a tree of sequents satisfying the following conditions:

- The topmost sequents of $\mathcal{D}$ are identity axioms.

- Every sequent in $\mathcal{D}$ except the lowest one is an upper sequent of an inference whose lower sequent is also in $\mathcal{D}$.

A sequent $\Gamma \Rightarrow \Delta$ is *deducible* if there is a proof having $\Gamma \Rightarrow \Delta$ as its the bottom sequent.

The Cut rule differs from the other rules in an important respect. With the rules for introduction of a connective on the left or the right, one sees that every formula that occurs above the line occurs below the line either directly, or as a subformula of a formula below the line, and that is also true for the structural rules. (Here $A(t)$ is counted as a subformula, in a slightly extended sense, of both $\exists x A(x)$ and $\forall x A(x)$.) But in the case of the Cut rule, the cut formula A vanishes. Gentzen showed that such "vanishing rules" can be eliminated.

Theorem 1.5 (Gentzen's Hauptsatz). If a sequent $\Gamma \Rightarrow \Delta$ is provable, then it is provable without use of the Cut Rule (called a *cut-free proof*).

The secret to Gentzen's Hauptsatz is the symmetry of left and right rules for the logical connectives. The proof of the cut elimination theorem is rather intricate as the process of removing cuts interferes with the structural rules. The possibility of contraction accounts for the high cost of eliminating cuts. Let $|\mathcal{D}|$ be the *height* of the deduction $\mathcal{D}$. Also, let $\mathrm{rank}(\mathcal{D})$ be *supremum* of the *lengths of cut formulae* occurring in $\mathcal{D}$. Turning $\mathcal{D}$ into a cut-free deduction of the same end sequent results, in the worst case, in a deduction of height $\mathcal{H}(\mathrm{rank}(\mathcal{D}), |\mathcal{D}|)$ where $\mathcal{H}(0, n) = n$ and $\mathcal{H}(k+1, n) = 4^{\mathcal{H}(k,n)}$, yielding hyper-exponential growth.

The *Hauptsatz* has an important corollary which explains its crucial role in obtaining consistency proofs.

Corollary 1.6 (The Subformula Property). *If a sequent $\Gamma \Rightarrow \Delta$ is provable, then it has a deduction all of whose formulae are subformulae of the formulae of Γ and Δ.*

Corollary 1.7. *A contradiction, i.e. the empty sequent $\emptyset \Rightarrow \emptyset$, is not deducible.*

Proof. According to the Hauptsatz, if the empty sequent were deducible it would have a deduction without cuts. In a cut-free deduction of the empty sequent only empty sequents can occur. But such a deduction does not exist. □

While mathematics is based on logic, it cannot be developed solely on the basis of *pure logic*. What is needed in addition are *axioms* that assert the *existence* of *mathematical objects* and their properties. Logic plus axioms gives rise to (formal) *theories* such as *Peano arithmetic* or the axioms of *Zermelo–Fraenkel set theory*. What happens when we try to apply the procedure of cut elimination to theories? Well, axioms are poisonous to this procedure. It breaks down because the symmetry of the sequent calculus is lost. In general, we cannot remove cuts from deductions in a theory T when the cut formula is an axiom of T. However, sometimes the axioms of a theory are of *bounded syntactic complexity*. Then the procedure applies

partially in that one can remove all cuts that exceed the complexity of the axioms of T. This gives rise to *partial cut elimination*. It is a very important tool in proof theory. For example, it works very well if the axioms of a theory can be presented as *atomic intuitionistic sequents* (also called *Horn clauses*), yielding the completeness of *Robinsons resolution method*. Partial cut elimination also pays off in the case of *fragments* of **PA** and set theory with *restricted induction schemes*, be it induction on natural numbers or sets. This method can be used to extract bounds from proofs of Π^0_2 statements in such fragments.

Full arithmetic (i.e. **PA**), though, does not even allow for partial cut elimination since the induction axioms have unbounded complexity. However, one can remove the obstacle to cut elimination in a drastic way by going *infinite*. The so-called *ω-rule* consists of the two types of *infinitary inferences*:

$$\frac{\Gamma \Rightarrow \Delta, F(0);\ \Gamma \Rightarrow \Delta, F(1);\ \ldots\ ; \Gamma \Rightarrow \Delta, F(n);\ \ldots}{\Gamma \Rightarrow \Delta, \forall x\, F(x)}\ \omega\mathrm{R}$$

$$\frac{F(0), \Gamma \Rightarrow \Delta;\ F(1), \Gamma \Rightarrow \Delta;\ \ldots\ ; F(n), \Gamma \Rightarrow \Delta;\ \ldots}{\exists x\, F(x), \Gamma \Rightarrow \Delta}\ \omega\mathrm{L}$$

The price to pay will be that deductions become infinite objects, i.e. *infinite well-founded trees*.

The sequent-style version of Peano arithmetic with the *ω-rule* will be termed $\mathbf{PA}_\omega$. $\mathbf{PA}_\omega$ has no use for free variables. Thus free variables are discarded and all *terms* will be closed. All formulae of this system are therefore closed, too. The *numerals* are the terms $\bar{n}$, where $\bar{0} = 0$ and $\overline{n+1} = S\bar{n}$. We shall identify $\bar{n}$ with the natural number n. All terms t of $\mathbf{PA}_\omega$ evaluate to a numeral $\bar{n}$.

$\mathbf{PA}_\omega$ has all the inference rules of the sequent calculus except for $\forall R$ and $\exists L$. In their stead, $\mathbf{PA}_\omega$ has the ωR and ωL inferences. The *Axioms* of $\mathbf{PA}_\omega$ are the following: (i) $\emptyset \Rightarrow A$ if A is a *true* atomic sentence; (ii) $B \Rightarrow \emptyset$ if B is a *false* atomic sentence; (iii) $F(s_1, \ldots, s_n) \Rightarrow F(t_1, \ldots, t_n)$ if $F(s_1, \ldots, s_n)$ is an atomic sentence and s_i and t_i evaluate to the same numeral.

With the aid of the ω-rule, each instance of the induction scheme becomes logically deducible, albeit the price to pay will be that the proof tree becomes infinite. To describe the cost of cut elimination for $\mathbf{PA}_\omega$, we introduce the measures of *height* and *cut rank* of a $\mathbf{PA}_\omega$ deduction $\mathcal{D}$. We will notate this by

$$\mathcal{D} \ \vert\!\vert^{\alpha}_{k}\ \Gamma \Rightarrow \Delta\,.$$

The above relation is defined inductively following the buildup of the deduction $\mathcal{D}$. For the *cut rank* we need the definition of the *length*, $|A|$ of a formula: $|A| = 0$ if A is atomic; $|\neg A_0| = |A_0| + 1$; $|A_0 \Box A_1| = \max(|A_0, A_1|) + 1$ where $\Box = \wedge, \vee, \rightarrow$; $|\exists x\, F(x)| = |\forall x\, F(x)| = |F(0)| + 1$.

Now suppose the last inference of $\mathcal{D}$ is of the form

$$\frac{\begin{array}{cccc}\mathcal{D}_0 & & \mathcal{D}_n & \\ \Gamma_0 \Rightarrow \Delta_0 & \cdots & \Gamma_n \Rightarrow \Delta_n & \ldots\quad n < \tau\end{array}}{\Gamma \Rightarrow \Delta}\ I$$

where $\tau = 1, 2, \omega$ and the $\mathcal{D}_n$ are the immediate subdeductions of $\mathcal{D}$. If

$$\mathcal{D}_n \vdash_k^{\alpha_n} \Gamma_n \Rightarrow \Delta_n$$

and $\alpha_n < \alpha$ for all $n < \tau$ then

$$\mathcal{D} \vdash_k^{\alpha} \Gamma \Rightarrow \Delta$$

providing that in the case of I being a *cut* with cut formula A we also have $|A| < k$. We will write $\mathbf{PA}_\omega \vdash_k^{\alpha} \Gamma \Rightarrow \Delta$ to convey that there exists a $\mathbf{PA}_\omega$-deduction $\mathcal{D} \vdash_k^{\alpha} \Gamma \Rightarrow \Delta$. The ordinal analysis of **PA** proceeds by first unfolding any **PA**-deduction into a $\mathbf{PA}_\omega$-deduction:

$$\text{If } \mathbf{PA} \vdash \Gamma \Rightarrow \Delta, \text{ then } \mathbf{PA}_\omega \vdash_k^{\omega+m} \Gamma \Rightarrow \Delta \tag{6}$$

for some $m, k < \omega$. The next step is to get rid of the cuts. It turns out that the cost of lowering the cut rank from $k+1$ to k is an exponential with base ω.

Theorem 1.8 (Cut Elimination for $\mathbf{PA}_\omega$).

$$\textit{If } \mathbf{PA}_\omega \vdash_{k+1}^{\alpha} \Gamma \Rightarrow \Delta, \textit{ then } \mathbf{PA}_\omega \vdash_k^{\omega^\alpha} \Gamma \Rightarrow \Delta.$$

As a result, if $\mathbf{PA}_\omega \vdash_n^{\alpha} \Gamma \Rightarrow \Delta$, we may apply the previous theorem n times to arrive at a cut-free deduction $\mathbf{PA}_\omega \vdash_0^{\rho} \Gamma \Rightarrow \Delta$ with $\rho = \omega^{\omega^{\cdot^{\cdot^{\omega^\alpha}}}}$, where the stack has height n. Combining this with the result from (6), it follows that every sequent $\Gamma \Rightarrow \Delta$ deducible in **PA** has a cut-free deduction in $\mathbf{PA}_\omega$ of length $< \varepsilon_0$. Ruminating on the details of how this result was achieved yields a consistency proof for **PA** from transfinite induction up to ε_0 for elementary decidable predicates on the basis of finitistic reasoning (as described in (1)).

Deductions in $\mathbf{PA}_\omega$ being well-founded infinite trees, they have a natural associated ordinal length, namely: the height of the tree as an ordinal. Thus the passage from finite deductions in **PA** to infinite cut-free deductions in $\mathbf{PA}_\omega$ provides an explanation of how the ordinal ε_0 is connected with **PA**.

Gentzen, however, did not consider infinite proof trees. The infinitary version of **PA** with the ω-rule was introduced by Schütte in [35]. Incidentally, the ω-rule had already been proposed by Hilbert [18]. Gentzen worked with finite deductions in the sequent calculus version of **PA**, devising an ingenious method of assigning ordinals to purported derivations of the empty sequent (inconsistency). It turns out in recent work by Buchholz [9] that in fact there is a much closer intrinsic connection between the way Gentzen assigned ordinals to deductions in **PA** and the way that ordinals are assigned to infinite deductions in $\mathbf{PA}_\omega$.

In the 1950s infinitary proof theory flourished in the hands of Schütte. He extended his approach to **PA** to systems of ramified analysis and brought this technique to perfection in his monograph "Beweistheorie" [36]. The ordinal representation systems necessary for Schütte's work will be reviewed in the next subsection.

1.3. A brief history of ordinal representation systems: 1904–1950. Ordinals assigned as lengths to deductions to keep track of the cost of operations such as cut elimination render ordinal analyses of theories particularly transparent. In the case of **PA**, Gentzen could rely on Cantor's normal form for a supply of ordinal representations. For stronger theories, though, segments larger than ε_0 have to be employed. Ordinal representation systems utilized by proof theorists in the 1960s arose in a purely set-theoretic context. This subsection will present some of the underlying ideas as progress in ordinal-theoretic proof theory also hinges on the development of sufficiently strong and transparent ordinal representation systems.

In 1904, Hardy [17] wanted to "construct" a subset of $\mathbb{R}$ of size $\aleph_1$. His method was to represent countable ordinals via increasing sequence of natural numbers and then to correlate a decimal expansion with each such sequence. Hardy used two processes on sequences: (i) Removing the first element to represent the successor; (ii) Diagonalizing at limits. E.g., if the sequence $1, 2, 3, \ldots$ represents the ordinal 1, then $2, 3, 4, \ldots$ represents the ordinal 2 and $3, 4, 5, \ldots$ represents the ordinal 3 etc., while the 'diagonal' $1, 3, 5, \ldots$ provides a representation of ω. In general, if $\lambda = \lim_{n\in\mathbb{N}} \lambda_n$ is a limit ordinal with $b_{n1}, b_{n2}, b_{n3}, \ldots$ representing $\lambda_n < \lambda$, then $b_{11}, b_{22}, b_{33}, \ldots$ represents λ. This representation, however, depends on the sequence chosen with limit λ. A sequence $(\lambda_n)_{n\in\mathbb{N}}$ with $\lambda_n < \lambda$ and $\lim_{n\in\mathbb{N}} \lambda_n = \lambda$ is called a *fundamental sequence* for λ. Hardy's two operations give explicit representations for all ordinals $< \omega^2$.

Veblen [44] extended the initial segment of the countable for which fundamental sequences can be given effectively. The new tools he devised were the operations of *derivation* and *transfinite iteration* applied to *continuous increasing functions* on ordinals.

Definition 1.9. Let ON be the class of ordinals. A (class) function $f : ON \to ON$ is said to be *increasing* if $\alpha < \beta$ implies $f(\alpha) < f(\beta)$ and *continuous* (in the order topology on ON) if

$$f(\lim_{\xi<\lambda} \alpha_\xi) = \lim_{\xi<\lambda} f(\alpha_\xi)$$

holds for every limit ordinal λ and increasing sequence $(\alpha_\xi)_{\xi<\lambda}$. f is called *normal* if it is increasing and continuous.

The function $\beta \mapsto \omega + \beta$ is normal while $\beta \mapsto \beta + \omega$ is not continuous at ω since $\lim_{\xi<\omega}(\xi + \omega) = \omega$ but $(\lim_{\xi<\omega} \xi) + \omega = \omega + \omega$.

Definition 1.10. The *derivative* f' of a function $f : ON \to ON$ is the function which enumerates in increasing order the solutions of the equation $f(\alpha) = \alpha$, also called the *fixed points* of f.

If f is a normal function, $\{\alpha : f(\alpha) = \alpha\}$ is a proper class and f' will be a normal function, too.

Definition 1.11. Now, given a normal function $f : ON \to ON$, define a hierarchy of normal functions as follows:

$$f_0 = f, \quad f_{\alpha+1} = f'_\alpha,$$
$$f_\lambda(\xi) = \xi^{th} \text{ element of } \bigcap_{\alpha<\lambda} (\text{Range of } f_\alpha) \text{ for } \lambda \text{ a limit ordinal.}$$

In this way, from the normal function f we get a two-place function, $\varphi_f(\alpha, \beta) := f_\alpha(\beta)$. Veblen then discusses the hierarchy when $f = \ell$, where $\ell(\alpha) = 1 + \alpha$.

The least ordinal $\gamma > 0$ closed under φ_ℓ, i.e. the least ordinal > 0 satisfying $(\forall \alpha, \beta < \gamma)\ \varphi_\ell(\alpha, \beta) < \gamma$ is the famous ordinal Γ_0 which Feferman [13] and Schütte [37], [38] determined to be the least ordinal 'unreachable' by *predicative means*.

Veblen extended this idea first to arbitrary finite numbers of arguments, but then also to transfinite numbers of arguments, with the proviso that in, for example $\Phi_f(\alpha_0, \alpha_1, \ldots, \alpha_\eta)$, only a finite number of the arguments α_ν may be non-zero. Finally, Veblen singled out the ordinal $E(0)$, where $E(0)$ is the least ordinal $\delta > 0$ which cannot be named in terms of functions $\Phi_\ell(\alpha_0, \alpha_1, \ldots, \alpha_\eta)$ with $\eta < \delta$, and each $\alpha_\gamma < \delta$.

Though the "great Veblen number" (as $E(0)$ is sometimes called) is quite an impressive ordinal it does not furnish an ordinal representation sufficient for the task of analyzing a theory as strong as Π^1_1 comprehension. Of course, it is possible to go beyond $E(0)$ and initiate a new hierarchy based on the function $\xi \mapsto E(\xi)$ or even consider hierarchies utilizing finite type functionals over the ordinals. Still all these further steps amount to rather mundane progress over Veblen's methods. In 1950 Bachmann [3] presented a new kind of operation on ordinals which dwarfs all hierarchies obtained by iterating Veblen's methods. Bachmann builds on Veblen's work but his novel idea was the systematic use of *uncountable ordinals* to keep track of the functions defined by diagonalization. Let Ω be the first uncountable ordinal. Bachmann defines a set of ordinals $\mathfrak{B}$ closed under successor such that with each limit $\lambda \in \mathfrak{B}$ is associated an increasing sequence $\langle \lambda[\xi] : \xi < \tau_\lambda \rangle$ of ordinals $\lambda[\xi] \in \mathfrak{B}$ of length $\tau_\lambda \in \mathfrak{B}$ and $\lim_{\xi<\tau_\lambda} \lambda[\xi] = \lambda$. A hierarchy of functions $(\varphi^{\mathfrak{B}}_\alpha)_{\alpha\in\mathfrak{B}}$ is then obtained as follows:

$$\begin{aligned}
&\varphi^{\mathfrak{B}}_0(\beta) = 1 + \beta, \qquad \varphi^{\mathfrak{B}}_{\alpha+1} = \left(\varphi^{\mathfrak{B}}_\alpha\right)', \\
&\varphi^{\mathfrak{B}}_\lambda \text{ enumerates } \bigcap_{\xi<\tau_\lambda} (\text{Range of } \varphi^{\mathfrak{B}}_{\lambda[\xi]}) \qquad \text{if } \lambda \text{ is a limit with } \tau_\lambda < \Omega, \\
&\varphi^{\mathfrak{B}}_\lambda \text{ enumerates } \{\beta < \Omega : \varphi^{\mathfrak{B}}_{\lambda[\beta]}(0) = \beta\} \qquad \text{if } \lambda \text{ is a limit with } \tau_\lambda = \Omega.
\end{aligned} \tag{7}$$

After the work of Bachmann, the story of ordinal representations becomes very complicated. Significant papers (by Isles, Bridge, Pfeiffer, Schütte, Gerber to mention a few) involve quite horrendous computations to keep track of the fundamental sequences. Also Bachmann's approach was combined with uses of higher type functionals by Aczel and Weyhrauch. Feferman proposed an entirely different method for

generating a Bachmann-type hierarchy of normal functions which does not involve fundamental sequences. Buchholz further simplified the systems and proved their recursivity. For details we recommend the preface to [7].

2. Ordinal analyses of systems of second order arithmetic and set theory

Ordinal analysis is concerned with theories serving as frameworks for formalising significant parts of mathematics. It is known that virtually all of ordinary mathematics can be formalized in Zermelo–Fraenkel set theory with the axiom of choice, **ZFC**. Hilbert and Bernays [19] showed that large chunks of mathematics can already be formalized in second order arithmetic. Owing to these observations, proof theory has been focusing on set theories and subsystems of second order arithmetic. Further scrutiny revealed that a small fragment is sufficient. Under the rubric of *Reverse Mathematics* a research programme has been initiated by Harvey Friedman some thirty years ago. The idea is to ask whether, given a theorem, one can prove its equivalence to some axiomatic system, with the aim of determining what proof-theoretical resources are necessary for the theorems of mathematics. More precisely, the objective of reverse mathematics is to investigate the role of set existence axioms in ordinary mathematics. The main question can be stated as follows:

> *Given a specific theorem τ of ordinary mathematics, which set existence axioms are needed in order to prove τ?*

Central to the above is the reference to what is called 'ordinary mathematics'. This concept, of course, doesn't have a precise definition. Roughly speaking, by ordinary mathematics we mean main-stream, non-set-theoretic mathematics, i.e. the core areas of mathematics which make no essential use of the concepts and methods of set theory and do not essentially depend on the theory of uncountable cardinal numbers.

2.1. Subsystems of second order arithmetic. The framework chosen for studying set existence in reverse mathematics, though, is second order arithmetic rather than set theory. Second order arithmetic, $\mathbf{Z_2}$, is a two-sorted formal system with one sort of variables $x, y, z, \ldots$ ranging over natural numbers and the other sort $X, Y, Z, \ldots$ ranging over sets of natural numbers. The language $\mathcal{L}_2$ of second-order arithmetic also contains the symbols of **PA**, and in addition has a binary relation symbol $\in$ for elementhood. Formulae are built from the prime formulae $s = t$, $s < t$, and $s \in X$ (where s, t are numerical terms, i.e. terms of **PA**) by closing off under the connectives $\wedge, \vee, \rightarrow, \neg$, numerical quantifiers $\forall x, \exists x$, and set quantifiers $\forall X, \exists X$.

The basic arithmetical axioms in all theories of second-order arithmetic are the defining axioms for $0, 1, +, \times, E, <$ (as for **PA**) and the *induction axiom*

$$\forall X(0 \in X \wedge \forall x(x \in X \rightarrow x + 1 \in X) \rightarrow \forall x(x \in X)).$$

We consider the axiom schema of $\mathcal{C}$-*comprehension* for formula classes $\mathcal{C}$ which is given by

$$\mathcal{C}\text{-}\mathbf{CA} \quad \exists X \forall u (u \in X \leftrightarrow F(u))$$

for all formulae $F \in \mathcal{C}$ in which X does not occur. Natural formula classes are the *arithmetical formulae*, consisting of all formulae without second order quantifiers $\forall X$ and $\exists X$, and the Π^1_n-*formulae*, where a Π^1_n-formula is a formula of the form $\forall X_1 \dots Q X_n \, A(X_1, \dots, X_n)$ with $\forall X_1 \dots Q X_n$ being a string of n alternating set quantifiers, commencing with a universal one, followed by an arithmetical formula $A(X_1, \dots, X_n)$.

For each axiom scheme $\mathbf{Ax}$ we denote by $(\mathbf{Ax})_0$ the theory consisting of the basic arithmetical axioms plus the scheme $\mathbf{Ax}$. By contrast, $(\mathbf{Ax})$ stands for the theory $(\mathbf{Ax})_0$ augmented by the scheme of induction for all $\mathcal{L}_2$-formulae.

An example for these notations is the theory $(\Pi^1_1\text{-}\mathbf{CA})_0$ which has the comprehension schema for Π^1_1-formulae.

In **PA** one can define an elementary injective pairing function on numbers, e.g $(n, m) := 2^n \times 3^m$. With the help of this function an infinite sequence of sets of natural numbers can be coded as a single set of natural numbers. The n^{th} section of set of natural numbers U is defined by $U_n := \{m : (n, m) \in U\}$. Using this coding, we can formulate the axiom of choice for formulae F in $\mathcal{C}$ by

$$\mathcal{C}\text{-}\mathbf{AC} \quad \forall x \exists Y \, F(x, Y) \to \exists Y \forall x \, F(x, Y_x).$$

For many mathematical theorems τ, there is a weakest natural subsystem $S(\tau)$ of $\mathbf{Z_2}$ such that $S(\tau)$ proves τ. Very often, if a theorem of ordinary mathematics is proved from the weakest possible set existence axioms, the statement of that theorem will turn out to be provably equivalent to those axioms over a still weaker base theory. This theme is referred to as *Reverse Mathematics*. Moreover, it has turned out that $S(\tau)$ often belongs to a small list of specific subsystems of $\mathbf{Z_2}$ dubbed $\mathbf{RCA}_0$, $\mathbf{WKL_0}$, $\mathbf{ACA}_0$, $\mathbf{ATR}_0$ and $(\Pi^1_1\text{-}\mathbf{CA})_0$, respectively. The systems are enumerated in increasing strength. The main set existence axioms of $\mathbf{RCA}_0$, $\mathbf{WKL_0}$, $\mathbf{ACA}_0$, $\mathbf{ATR}_0$, and $(\Pi^1_1\text{-}\mathbf{CA})_0$ are recursive comprehension, weak König's lemma, arithmetical comprehension, arithmetical transfinite recursion, and Π^1_1-comprehension, respectively. For exact definitions of all these systems and their role in reverse mathematics see [40]. The proof-theoretic strength of $\mathbf{RCA}_0$ is weaker than that of **PA** while $\mathbf{ACA}_0$ has the same strength as **PA**. Let $|T| = |T|_{\mathsf{Con}}$. To get a sense of scale, the strengths of the first four theories are best expressed via their proof-theoretic ordinals: $|\mathbf{RCA}_0| = |\mathbf{WKL_0}| = \omega^\omega$, $|\mathbf{ACA}_0| = \varepsilon_0$, $|\mathbf{ATR}_0| = \Gamma_0$. $|(\Pi^1_1\text{-}\mathbf{CA})_0|$, however, eludes expression in the ordinal representations introduced so far. $\Pi^1_1\text{-}\mathbf{CA}$ involves a so-called *impredicative definition*. An impredicative definition of an object refers to a presumed totality of which the object being defined is itself to be a member. For example, to define a set of natural numbers X as $X = \{n \in \mathbb{N} : \forall Y \subseteq \mathbb{N} \, F(n, Y)\}$ is impredicative since it involves the quantified variable 'Y' ranging over arbitrary subsets of the natural numbers $\mathbb{N}$, of which the set X being defined is one member.

Determining whether $\forall Y \subseteq \mathbb{N}\ F(n, Y)$ holds involves an apparent circle since we shall have to know in particular whether $F(n, X)$ holds – but that cannot be settled until X itself is determined. Impredicative set definitions permeate the fabric of Zermelo–Fraenkel set theory in the guise of the separation and replacement axioms as well as the powerset axiom.

A major breakthrough was made by Takeuti in 1967, who for the first time obtained an ordinal analysis of an impredicative theory. In [41] he gave an ordinal analysis of (Π^1_1-**CA**), extended in 1973 to (Π^1_1-**AC**) in [43] jointly with Yasugi. For this Takeuti returned to Gentzen's method of assigning ordinals (ordinal diagrams, to be precise) to purported derivations of the empty sequent (inconsistency).

The next wave of results, which concerned theories of iterated inductive definitions, were obtained by Buchholz, Pohlers, and Sieg in the late 1970s (see [7]). Takeuti's methods of reducing derivations of the empty sequent ("the inconsistency") were extremely difficult to follow, and therefore a more perspicuous treatment was to be hoped for. Since the use of the infinitary ω-rule had greatly facilitated the ordinal analysis of number theory, new infinitary rules were sought. In 1977 (see [5]) Buchholz introduced such rules, dubbed Ω-rules to stress the analogy. They led to a proof-theoretic treatment of a wide variety of systems, as exemplified in the monograph [8] by Buchholz and Schütte. Yet simpler infinitary rules were put forward a few years later by Pohlers, leading to the *method of local predicativity*, which proved to be a very versatile tool (see [23]).

2.2. Set theories. With the work of Jäger and Pohlers (see [20], [21]) the forum of ordinal analysis then switched from the realm of second-order arithmetic to set theory, shaping what is now called *admissible proof theory*, after the models of *Kripke–Platek set theory*, **KP**. Their work culminated in the analysis of the system Π^1_1-**AC** plus an induction principle called *Bar Induction* **BI** which is a scheme asserting that transfinite induction along well-founded relations holds for arbitrary formulae (see [21]).

By and large, ordinal analyses for set theories are more uniform and transparent than for subsystems of $\mathbf{Z_2}$. The axiom systems for set theories considered in this paper are formulated in the usual language of set theory (called $\mathcal{L}_\in$ hereafter) containing $\in$ as the only non-logical symbol besides $=$. Formulae are built from prime formulae $a \in b$ and $a = b$ by use of propositional connectives and quantifiers $\forall x, \exists x$. Quantifiers of the forms $\forall x \in a$, $\exists x \in a$ are called *bounded*. *Bounded* or Δ_0*-formulae* are the formulae wherein all quantifiers are bounded; Σ_1-formulae are those of the form $\exists x \varphi(x)$ where $\varphi(a)$ is a Δ_0-formula. For $n > 0$, Π_n-formulae (Σ_n-formulae) are the formulae with a prefix of n alternating unbounded quantifiers starting with a universal (existential) one followed by a Δ_0-formula. The class of Σ-formulae is the smallest class of formulae containing the Δ_0-formulae which is closed under $\wedge$, $\vee$, bounded quantification and unbounded existential quantification.

One of the set theories which is amenable to ordinal analysis is Kripke–Platek set theory, **KP**. Its standard models are called *admissible sets*. One of the reasons that this is an important theory is that a great deal of set theory requires only the

axioms of **KP**. An even more important reason is that admissible sets have been a major source of interaction between model theory, recursion theory and set theory (cf. [4]). **KP** arises from **ZF** by completely omitting the power set axiom and restricting separation and collection to bounded formulae. These alterations are suggested by the informal notion of 'predicative'. To be more precise, the axioms of **KP** consist of *Extensionality, Pair, Union, Infinity, Bounded Separation*

$$\exists x\, \forall u\, [u \in x \leftrightarrow (u \in a \ \wedge\ F(u))]$$

for all bounded formulae $F(u)$, *Bounded Collection*

$$\forall x \in a\, \exists y\, G(x, y) \ \rightarrow\ \exists z\, \forall x \in a\, \exists y \in z\, G(x, y)$$

for all bounded formulae $G(x, y)$, and *Set Induction*

$$\forall x\ [(\forall y \in x\ H(y)) \rightarrow H(x)] \ \rightarrow\ \forall x\, H(x)$$

for all formulae $H(x)$.

A transitive set A such that $(A, \in)$ is a model of **KP** is called an *admissible set*. Of particular interest are the models of **KP** formed by segments of Gödel's *constructible hierarchy* **L**. The constructible hierarchy is obtained by iterating the definable powerset operation through the ordinals

$$\begin{aligned}
\mathbf{L}_0 &= \emptyset,\\
\mathbf{L}_\lambda &= \bigcup\{\mathbf{L}_\beta : \beta < \lambda\}\ \lambda \text{ limit}\\
\mathbf{L}_{\beta+1} &= \big\{X :\ X \subseteq \mathbf{L}_\beta;\ X \text{ definable over } \langle \mathbf{L}_\beta, \in\rangle\big\}.
\end{aligned}$$

So any element of **L** of level α is definable from elements of **L** with levels $< \alpha$ and the parameter $\mathbf{L}_\alpha$. An ordinal α is *admissible* if the structure $(\mathbf{L}_\alpha, \in)$ is a model of **KP**.

Formulae of $\mathcal{L}_2$ can be easily translated into the language of set theory. Some of the subtheories of $\mathbf{Z_2}$ considered above have set-theoretic counterparts, characterized by extensions of **KP**. **KPi** is an extension of **KP** via the axiom

$$(Lim) \quad \forall x \exists y [x \in y \ \wedge\ y \textit{ is an admissible set}].$$

KPl denotes the system **KPi** without Bounded Collection. It turns out that $(\Pi^1_1\text{-}\mathbf{AC})+\mathbf{BI}$ proves the same $\mathcal{L}_2$-formulae as **KPi**, while $(\Pi^1_1\text{-}\mathbf{CA})$ proves the same $\mathcal{L}_2$-formulae as **KPl**.

2.3. Sketches of an ordinal analysis of KP. Serving as a miniature example of an ordinal analysis of an impredicative system, the ordinal analysis of **KP** (see [20], [6]) we will sketched in broad strokes. Bachmann's system can be recast without fundamental sequences as follows: Let Ω be a "big" ordinal, e.g. $\Omega = \aleph_1$. By

recursion on α we define sets $C^{\Omega}(\alpha, \beta)$ and the ordinal $\psi_{\Omega}(\alpha)$ as follows:

$$C^{\Omega}(\alpha, \beta) = \begin{cases} \text{closure of } \beta \cup \{0, \Omega\} \text{ under:} \\ +, (\xi \mapsto \omega^{\xi}) \\ (\xi \longmapsto \psi_{\Omega}(\xi))_{\xi < \alpha} \end{cases} \tag{8}$$

$$\psi_{\Omega}(\alpha) \simeq \min\{\rho < \Omega : \ C^{\Omega}(\alpha, \rho) \cap \Omega = \rho \}. \tag{9}$$

It can be shown that $\psi_{\Omega}(\alpha)$ is always defined and that $\psi_{\Omega}(\alpha) < \Omega$. Moreover, $[\psi_{\Omega}(\alpha), \Omega) \ \cap \ C^{\Omega}(\alpha, \psi_{\Omega}(\alpha)) = \emptyset$; thus the order-type of the ordinals below Ω which belong to the set $C^{\Omega}(\alpha, \psi_{\Omega}(\alpha))$ is $\psi_{\Omega}(\alpha)$. $\psi_{\Omega}(\alpha)$ is also a countable ordinal. In more pictorial terms, $\psi_{\Omega}(\alpha)$ is the α^{th} *collapse* of Ω.

Let $\varepsilon_{\Omega+1}$ be the least ordinal $\alpha > \Omega$ such that $\omega^{\alpha} = \alpha$. The set of ordinals $C^{\Omega}(\varepsilon_{\Omega+1}, 0)$ gives rise to an elementary computable ordinal representation system. In what follows, $C^{\Omega}(\varepsilon_{\Omega+1}, 0)$ will be abbreviated to $\mathcal{T}(\Omega)$.

In the case of **PA** the addition of an infinitary rule restored the possibility of cut elimination. In order to obtain a similar result for set theories like **KP**, one has to work a bit harder. A peculiarity of **PA** is that every object n of the intended model has a canonical name in the language, namely, the n^{th} numeral. It is not clear, though, how to bestow a canonical name to each element of the set-theoretic universe. This is where *Gödel's constructible universe* **L** comes in handy. As **L** is "made" from the ordinals it is pretty obvious how to "name" sets in **L** once one has names for ordinals. These will be taken from $\mathcal{T}(\Omega)$. Henceforth, we shall restrict ourselves to ordinals from $\mathcal{T}(\Omega)$. The set terms and their ordinal levels are defined inductively. First, for each $\alpha \in \mathcal{T}(\Omega) \cap \Omega$, there will be a set term $\mathbb{L}_{\alpha}$. Its ordinal level is declared to be α. If $F(a, \vec{b})$ is a set-theoretic formula (whose free variables are among the indicated) and $\vec{s} \equiv s_1, \ldots, s_n$ are set terms with levels $< \alpha$, then the formal expression $\{x \in \mathbb{L}_{\alpha} : F(x, \vec{s})^{\mathbb{L}_{\alpha}}\}$ is a set term of level α. Here $F(x, \vec{s})^{\mathbb{L}_{\alpha}}$ results from $F(x, \vec{s})$ by restricting all unbounded quantifiers to $\mathbb{L}_{\alpha}$.

The collection of set terms will serve as a formal universe for a theory $\mathbf{KP}_{\infty}$ with infinitary rules. The infinitary rule for the universal quantifier on the right takes the form: From $\Gamma \Rightarrow \Delta, F(t)$ for all RS_{Ω}-terms t conclude $\Gamma \Rightarrow \Delta, \forall x \ F(x)$. There are also rules for bounded universal quantifiers: From $\Gamma \Rightarrow \Delta, F(t)$ for all RS_{Ω}-terms t with levels $< \alpha$ conclude $\Gamma \Rightarrow \Delta, (\forall x \in \mathbb{L}_{\alpha}) \ F(x)$. The corresponding rule for introducing a universal quantifier bounded by a term of the form $\{x \in \mathbb{L}_{\alpha} : F(x, \vec{s})^{\mathbb{L}_{\alpha}}\}$ is slightly more complicated. With the help of these infinitary rules it now possible to give logical deductions of all axioms of **KP** with the exception of Bounded Collection. The latter can be deduced from the rule of Σ-Reflection: From $\Gamma \Rightarrow \Delta, C$ conclude $\Gamma \Rightarrow \Delta, \exists z \, C^z$ for every Σ-formula C. The class of Σ-formulae is the smallest class of formulae containing the bounded formulae which is closed under $\wedge$, $\vee$, bounded quantification and unbounded existential quantification. C^z is obtained from C by replacing all unbounded quantifiers $\exists x$ in C by $\exists x \in z$.

The length and cut ranks of $\mathbf{KP}_\infty$-deductions will be measured by ordinals from $\mathcal{T}(\Omega)$. If

$$\mathbf{KP} \vdash F(u_1, \ldots, u_r)$$

then $\mathbf{KP}_\infty \vdash^{\Omega\cdot m}_{\Omega+n} B(s_1, \ldots, s_r)$ holds for some m, n and all set terms $s_1, \ldots, s_r$; m and n depend only on the **KP**-derivation of $B(\vec{u})$.

The usual cut elimination procedure works unless the cut formulae have been introduced by Σ-reflection rules. The obstacle to pushing cut elimination further is exemplified by the following scenario:

$$\dfrac{\dfrac{\vdash^{\delta}_{\Omega} \Gamma \Rightarrow \Delta, C}{\vdash^{\xi}_{\Omega} \Gamma \Rightarrow \Delta, \exists z\, C^z}(\Sigma\text{-Ref}) \qquad \dfrac{\cdots \vdash^{\xi_s}_{\Omega} \Xi, C^s \Rightarrow \Lambda \cdots (|s| < \Omega)}{\vdash^{\xi}_{\Omega} \Xi, \exists z\, C^z \Rightarrow \Lambda}(\exists L)}{\vdash^{\alpha}_{\Omega+1} \Gamma, \Xi \Rightarrow \Delta, \Lambda}(\text{Cut})$$

In general, it won't be possible to remove such an instance of the Cut Rule. However, if the complexity of the side formulae is just right, the cut can be removed by a technique called *collapsing of deductions*. This method applies when the formulae in Γ and Ξ are Π-formulae and the formulae in Δ and Λ are Σ-formulae. The class of Π-formulae is the smallest class of formulae containing the bounded formulae which is closed under $\wedge$, $\vee$, bounded quantification and unbounded universal quantification.

For the technique of collapsing one needs the function $\alpha \mapsto \psi_\Omega(\alpha)$ and, moreover, it is necessary to ensure that the infinite deductions are of a very uniform character. The details are rather finicky and took several years to work out. The upshot is that every Σ sentence C deducible in **KP** has a cut-free deduction in $\mathbf{KP}_\infty$ of length $\psi_\Omega(\varepsilon_{\Omega+1})$, which entails that $L_{\psi_\Omega(\varepsilon_{\Omega+1})} \models C$. Moreover, the proof-theoretic ordinal of **KP** is $\psi_\Omega(\varepsilon_{\Omega+1})$, also known as the *Bachmann–Howard ordinal*.

2.4. Admissible proof theory. **KP** is the weakest in a line of theories that were analyzed by proof theorists of the Munich school in the late 1970s and 1980s. In many respects, **KP** is a very special case. Several fascinating aspects of ordinal analysis do not yet exhibit themselves at the level of **KP**.

Recall that **KPl** is the set-theoretic version of $(\Pi^1_1\text{-}\mathbf{AC}) + \mathbf{BI}$, while **KPi** is the set-theoretic counterpart to $(\Pi^1_1\text{-}\mathbf{AC}) + \mathbf{BI}$. The main axiom of **KPl** says that every set is contained in an admissible set (one also says that the admissible sets are cofinal in the universe) without requiring that the universe is also admissible, too. To get a sense of scale for comparing **KP**, **KPl**, and **KPi** it is perhaps best to relate the large cardinal assumptions that give rise to the pertaining ordinal representation systems. In the case of **KPl** the assumptions is that there are infinitely many large ordinals Ω_1, Ω_2, Ω_3, ... (where Ω_n can be taken to be $\aleph_n$) each equipped with their own 'collapsing' function $\alpha \mapsto \psi_{\Omega_n}(\alpha)$. The ordinal system sufficient for **KPi** is built using the much bolder assumption that there is an inaccessible cardinal I.

As the above set theories are based on the notion of admissible set it is suitable to call the proof theory concerned with them 'admissible proof theory'. The salient

feature of admissible sets is that they are models of Bounded Collection and that that principle is equivalent to Σ Reflection on the basis of the other axioms of **KP** (see [4]). Furthermore, admissible sets of the form $\mathbf{L}_\kappa$ also satisfy Π_2 reflection, i.e., if $\mathbf{L}_\kappa \models \forall x\, \exists y\, C(x, y, \vec{a})$ with $C(x, y)$ bounded and $\vec{a} \in \mathbf{L}_\kappa$, then there exists $\rho < \kappa$ such that $\vec{a} \in \mathbf{L}_\rho$ and $\mathbf{L}_\rho \models \forall x\, \exists y\, C(x, y, \vec{a})$.

In essence, admissible proof theory is a gathering of cut-elimination and collapsing techniques that can handle infinitary calculi of set theory with Σ and/or Π_2 reflection rules, and thus lends itself to ordinal analyses of theories of the form **KP**+ *"there are x many admissibles"* or **KP**+ *"there are many admissibles"*.

A theory on the verge of admissible proof theory is **KPM**, designed to axiomatize essential features of a recursively Mahlo universe of sets. An admissible ordinal κ is said to be recursively Mahlo if it satisfies Π_2-reflection in the above sense but with the extra condition that the reflecting set $\mathbf{L}_\rho$ be admissible as well. The ordinal representation [25] for **KPM** is built on the assumption that there exists a Mahlo cardinal. The novel feature of over previous work is that there are two layers of collapsing functions. The ordinal analysis for **KPM** was carried out in [26]. A different approach to **KPM** using ordinal diagrams is due to Arai [1].

The means of admissible proof theory are too weak to deal with the next level of reflection having three alternations of quantifiers, i.e. Π_3-reflection.

2.5. Rewards of ordinal analysis Results that have been achieved through ordinal analysis mainly fall into four groups: (1) Consistency of subsystems of classical second order arithmetic and set theory relative to constructive theories, (2) reductions of theories formulated as conservation theorems, (3) combinatorial independence results, and (4) classifications of provable functions and ordinals. A detailed account of these results has been given in [31], section 3. An example where ordinal representation systems led to a new combinatorial result was Friedman's extension of Kruskal's Theorem, EKT, which asserts that finite trees are well-quasi-ordered under gap embeddability (see [39]). The gap condition imposed on the embeddings is directly related to an ordinal notation system that was used for the analysis of Π^1_1 comprehension. The principle EKT played a crucial role in the proof of the graph minor theorem of Robertson and Seymour (see [16]).

Theorem 2.1 (Robertson, Seymour). *For any infinite sequence $G_0, G_1, G_2, \ldots$ of finite graphs there exist $i < j$ so that G_i is isomorphic to a minor of G_j.*

3. Beyond admissible proof theory

Gentzen fostered hopes that with sufficiently large constructive ordinals one could establish the consistency of analysis, i.e., $\mathbf{Z_2}$. The purpose of this section is to report on the next major step in analyzing fragments of $\mathbf{Z_2}$. This is obviously the ordinal

analysis of the system (Π^1_2-**CA**).[1] The strength of (Π^1_2-**CA**) dwarfs that of (Π^1_1-**AC**). The treatment of Π^1_2 comprehension posed formidable technical challenges (see [30], [32], [33]). Other approaches to ordinal analysis of systems above Π^1_1-**AC** are due to Arai (see [1], [2]) who uses ordinal diagrams and finite deductions, and Carlson [11] who employs patterns of resemblance.

In the following, we will gradually slice Π^1_2 comprehension into degrees of reflection to achieve a sense of scale. There is no way to describe this comprehension simply in terms of admissibility except that on the set-theoretic side, Π^1_2 comprehension corresponds to Σ_1 separation, i.e. the scheme of axioms

$$\exists z(z = \{x \in a : \phi(x)\})$$

for all Σ_1 formulas ϕ. The precise relationship is as follows:

Theorem 3.1. **KP** + Σ_1 *separation and* (Π^1_2-**CA**) + **BI** *prove the same sentences of second order arithmetic.*

The ordinals κ such that $\mathbf{L}_\kappa \models$ **KP** + Σ_1-*Separation* are familiar from ordinal recursion theory.

Definition 3.2. An admissible ordinal κ is said to be *nonprojectible* if there is no total κ-recursive function mapping κ one-one into some $\beta < \kappa$, where a function $g\colon \mathbf{L}_\kappa \to \mathbf{L}_\kappa$ is called *κ-recursive* if it is Σ definable in $\mathbf{L}_\kappa$.

The key to the 'largeness' properties of nonprojectible ordinals is that for any nonprojectible ordinal κ, $\mathbf{L}_\kappa$ is a limit of Σ_1-elementary substructures, i.e. for every $\beta < \kappa$ there exists a $\beta < \rho < \kappa$ such that $\mathbf{L}_\rho$ is a Σ_1-elementary substructure of $\mathbf{L}_\kappa$, written $\mathbf{L}_\rho \prec_1 \mathbf{L}_\kappa$.

Such ordinals satisfying $\mathbf{L}_\rho \prec_1 \mathbf{L}_\kappa$ have strong reflecting properties. For instance, if $\mathbf{L}_\rho \models C$ for some set-theoretic sentence C (containing parameters from $\mathbf{L}_\rho$), then there exists a $\gamma < \rho$ such that $\mathbf{L}_\gamma \models C$. This is because $\mathbf{L}_\rho \models C$ implies $\mathbf{L}_\kappa \models \exists \gamma\, C^{\mathbf{L}_\gamma}$, hence $\mathbf{L}_\rho \models \exists \gamma\, C^{\mathbf{L}_\gamma}$ using $\mathbf{L}_\rho \prec_1 \mathbf{L}_\kappa$.

The last result makes it clear that an ordinal analysis of Π^1_2 comprehension would necessarily involve a proof-theoretic treatment of reflections beyond those surfacing in admissible proof theory. The notion of stability will be instrumental.

Definition 3.3. α is *δ-stable* if $\mathbf{L}_\alpha \prec_1 \mathbf{L}_{\alpha+\delta}$.

For our purposes we need refinements of this notion, the simplest being provided by:

Definition 3.4. $\alpha > 0$ *is said to be* *Π_n-reflecting* if $\mathbf{L}_\alpha \models \Pi_n$-reflection. By Π_n-reflection we mean the scheme $C \to \exists z[\,\mathrm{Tran}(z) \wedge z \neq \emptyset \wedge C^z]$, where C is Π_n, and $\mathrm{Tran}(z)$ expresses that z is a transitive set.

[1] For more background information see [42], p. 259, [15], p. 362, [24], p. 374.

Π_n-reflection for all n suffices to express one step in the $\prec_1$ relation.

Lemma 3.5 (cf. [34], 1.18). $\mathbf{L}_\kappa \prec_1 \mathbf{L}_{\kappa+1}$ iff κ is Π_n-reflecting for all n.

The step of analyzing Kripke–Platek set theory augmented by Π_n-reflection rules was taken in [29]; the ordinal representation system for Π_3-reflection employed a weakly compact cardinal.

A further refinement of the notion of δ-stability will be addressed next.

Definition 3.6. κ is said to be δ-Π_n-reflecting if whenever $C(u, \vec{x})$ is a set-theoretic Π_n formula, $a_1, \ldots, a_r \in \mathbf{L}_\kappa$ and $\mathbf{L}_{\kappa+\delta} \models C[\kappa, a_1, \ldots, a_n]$, then there exists $\kappa_0, \delta_0 < \kappa$ such that $a_1, \ldots, a_r \in \mathbf{L}_{\kappa_0}$ and $\mathbf{L}_{\kappa_0+\delta_0} \models C[\kappa_0, a_1, \ldots, a_n]$.

Putting the previous definition to work, one gets:

Corollary 3.7. If κ is $\delta + 1$-Σ_1-reflecting, then, for all n, κ is δ-Σ_n-reflecting.

At this point let us return to proof theory to explain the need for even further refinements of the preceding notions. Recall that the first nonprojectible ordinal ρ is a limit of smaller ordinals ρ_n such that $\mathbf{L}_{\rho_n} \prec_1 \mathbf{L}_\rho$. In the ordinal representation system $\mathcal{OR}$ for Π^1_2-**CA**, there will be symbols $\mathfrak{E}_n$ and $\mathfrak{E}_\omega$ for ρ_n and ρ, respectively. The associated infinitary proof system will have rules

$$(\mathrm{Ref}_{\Sigma(\mathbb{L}_{\mathfrak{E}_n+\delta})}) \quad \frac{\Gamma \Rightarrow \Delta, C(\vec{s})^{\mathbb{L}_{\mathfrak{E}_n+\delta}}}{\Gamma \Rightarrow \Delta, (\exists z \in \mathbb{L}_{\mathfrak{E}_n})(\exists \vec{x} \in \mathbb{L}_{\mathfrak{E}_n})[\mathrm{Tran}(z) \wedge C(\vec{x})^z]},$$

where $C(\vec{x})$ is a Σ formula, $\vec{s}$ are set terms of levels $< \mathfrak{E}_n + \delta$, and $\delta < \mathfrak{E}_\omega$. These rules suffice to bring about the embedding $\mathbf{KP} + \Sigma_1$-Separation into the infinitary proof system, but reflection rules galore will be needed to carry out cut-elimination. For example, there will be "many" ordinals $\pi, \delta \in \mathcal{OR}$ that play the role of δ-Π_{n+1}-reflecting ordinals by virtue of corresponding reflection rules in the infinitary calculus.

4. A large cardinal notion

An important part of ordinal analysis is the development of ordinal representation systems. Extensive ordinal representation systems are difficult to understand from a purely syntactical point of view, often to such an extent that it makes no sense to present an ordinal representation system without giving some kind of semantic interpretation. Large cardinals have been used quite frequently in the definition procedure of strong ordinal representation systems, and large cardinal notions have been an important source of inspiration. In the end, they can be dispensed with, but they add an intriguing twist to the relation between set theory and proof theory. The advantage of working in a strong set-theoretic context is that we can build models without getting buried under complexity considerations.

Such systems are usually generated from collapsing functions. However, from now on we prefer to call them *projection functions* since they will no longer bear any resemblance to Mostowski's collapsing function. In [33], the projection functions needed for the ordinal analysis of Π^1_2 have been construed as inverses to certain partial elementary embeddings. In this final section we shall indicate a model for the projection functions, employing rather sweeping large cardinal axioms, in that we shall presume the existence of certain cardinals, featuring a strong form of indescribability, dubbed *shrewdness*.

To be able to eliminate reflections of the type described in Definition 3.6 requires projection functions which can project intervals $[\kappa, \kappa + \delta]$ of ordinals down below κ.

Definition 4.1. Let $V = \bigcup_{\alpha \in ON} V_\alpha$ be the cumulative hierarchy of sets, i.e.

$$V_0 = \emptyset, \quad V_{\alpha+1} = \{X : X \subseteq V_\alpha\}, \quad V_\lambda = \bigcup_{\xi<\lambda} V_\xi \text{ for limit ordinals } \lambda.$$

Let $\eta > 0$. A cardinal κ is *η-shrewd* if for all $P \subseteq V_\kappa$ and every set-theoretic formula $F(v_0, v_1)$, whenever

$$V_{\kappa+\eta} \models F[P, \kappa],$$

then there exist $0 < \kappa_0, \eta_0 < \kappa$ such that

$$V_{\kappa_0+\eta_0} \models F[P \cap V_{\kappa_0}, \kappa_0].$$

κ is *shrewd* if κ is η-shrewd for every $\eta > 0$.

Let $\mathcal{F}$ be a collection of formulae. A cardinal κ is *η-$\mathcal{F}$-shrewd* if for all $P \subseteq V_\kappa$ and every $\mathcal{F}$-formula $H(v_0, v_1)$, whenever

$$V_{\kappa+\eta} \models H[P, \kappa],$$

then there exist $0 < \kappa_0, \eta_0 < \kappa$ such that

$$V_{\kappa_0+\eta_0} \models H[P \cap V_{\kappa_0}, \kappa_0].$$

We will also consider a notion of shrewdness with regard to a given class.

Let $\mathbf{U}$ be a fresh unary predicate symbol. Given a language $\mathcal{L}$ let $\mathcal{L}(\mathbf{U})$ denote its extension by $\mathbf{U}$. If $\mathcal{A}$ is a class we denote by $\langle V_\alpha; \mathcal{A}\rangle$ the structure $\langle V_\alpha; \in; \mathcal{A} \cap V_\alpha\rangle$.

For an $\mathcal{L}_{\text{set}}(\mathbf{U})$-sentence ϕ, let the meaning of "$\langle V_\alpha; \mathcal{A}\rangle \models \phi$" be determined by interpreting $\mathbf{U}(t)$ as $t \in \mathcal{A} \cap V_\alpha$.

Definition 4.2. Assume that $\mathcal{A}$ is a class. Let $\eta > 0$. A cardinal κ is *$\mathcal{A}$-η-shrewd* if for all $P \subseteq V_\kappa$ and every formula $F(v_0, v_1)$ of $\mathcal{L}_{\text{set}}(\mathbf{U})$, whenever

$$\langle V_{\kappa+\eta}; \mathcal{A}\rangle \models F[P, \kappa],$$

then there exist $0 < \kappa_0, \eta_0 < \kappa$ such that

$$\langle V_{\kappa_0+\eta_0}; \mathcal{A}\rangle \models F[P \cap V_{\kappa_0}, \kappa_0].$$

κ is $\mathcal{A}$-*shrewd* if κ is $\mathcal{A}$-η-shrewd for every $\eta > 0$.

Likewise, for $\mathcal{F}$ a collection of formulae in a language $\mathcal{L}(\mathbf{U})$, we say that a cardinal κ is $\mathcal{A}$-η-$\mathcal{F}$-*shrewd* if for all $P \subseteq V_\kappa$ and every $\mathcal{F}$-formula $H(v_0, v_1)$, whenever

$$\langle V_{\kappa+\eta}; \mathcal{A}\rangle \models H[P, \kappa],$$

then there exist $0 < \kappa_0, \eta_0 < \kappa$ such that

$$\langle V_{\kappa_0+\eta_0}; \mathcal{A}\rangle \models H[P \cap V_{\kappa_0}, \kappa_0].$$

Corollary 4.3. If κ is $\mathcal{A}$-δ-shrewd and $0 < \eta < \delta$, then κ is $\mathcal{A}$-η-shrewd.

There are similarities between the notions of η-shrewdness and η-indescribability (see [12], Ch. 9, §4). However, it should be noted that if κ is η-indescribable and $\rho < \eta$, it does not necessarily follow that κ is also ρ-indescribable (see [12], 9.4.6).

A reason for calling the above cardinals *shrewd* is that if there is a shrewd cardinal κ in the universe, then, loosely speaking, for any notion of large cardinal N which does not make reference to the totality of all ordinals, if there exists an N-cardinal then the least such cardinal is below κ. So for instance, if there are measurable and shrewd cardinals in the universe, then the least measurable is smaller than the least shrewd cardinal.

To situate the notion of shrewdness with regard to consistency strength in the usual hierarchy of large cardinals, we recall the notion of a subtle cardinal.

Definition 4.4. A cardinal κ is said to be *subtle* if for any sequence $\langle S_\alpha : \alpha < \kappa\rangle$ such that $S_\alpha \subseteq \alpha$ and C closed and unbounded in κ, there are $\beta < \delta$ both in C satisfying

$$S_\delta \cap \beta = S_\beta.$$

Since subtle cardinals are not covered in many of the standard texts dealing with large cardinals, we mention the following facts (see [22], §20):

Remark 4.5. Let $\kappa(\omega)$ denote the first ω-Erdös cardinal.

(i) $\{\pi < \kappa(\omega) : \pi \text{ is subtle}\}$ is stationary in $\kappa(\omega)$.

(ii) 'Subtlety' relativises to $\mathbf{L}$, i.e. if π is subtle, then $\mathbf{L} \models$ "π is subtle".

Lemma 4.6. Assume that π is a subtle cardinal and that $\mathcal{A} \subseteq V_\pi$. Then for every $B \subseteq \pi$ closed and unbounded in π there exists $\kappa \in B$ such that

$$\langle V_\pi; \mathcal{A}\rangle \models \text{"}\kappa \text{ is } \mathcal{A}\text{-shrewd"}.$$

Corollary 4.7. Assume that π is a subtle cardinal. Then there exists a cardinal $\kappa < \pi$ such that κ is η-shrewd for all $\eta < \pi$.

References

[1] Arai, T., Proof theory for theories of ordinals I: recursively Mahlo ordinals. *Ann. Pure Appl. Logic* **122** (2003), 1–85.

[2] Arai, T., Proof theory for theories of ordinals II: Π_3-Reflection. *Ann. Pure Appl. Logic* **129** (2004), 39–92.

[3] Bachmann, H., Die Normalfunktionen und das Problem der ausgezeichneten Folgen von Ordinalzahlen. *Vierteljschr. Naturforsch. Ges. Zürich* **95** (1950), 115–147.

[4] Barwise, J., *Admissible Sets and Structures*. Perspectives in Mathematical Logic, Springer-Verlag, Berlin 1975.

[5] Buchholz, W., Eine Erweiterung der Schnitteliminationsmethode. Habilitationsschrift, München 1977.

[6] Buchholz, W., A simplified version of local predicativity. In *Proof theory* (Leeds, 1990), ed. by P. Aczel, H. Simmons, S. S. Wainer, Cambridge University Press, Cambridge 1993, 115–147.

[7] Buchholz, W., Feferman, S., Pohlers, W., Sieg, W., *Iterated inductive definitions and subsystems of analysis*. Lecture Notes in Math. 897, Springer-Verlag, Berlin 1981.

[8] Buchholz, W., and Schütte, K., *Proof theory of impredicative subsystems of analysis.* Stud. Proof Theory Monogr. 2, Bibliopolis, Naples 1988.

[9] Buchholz, W., Explaining Gentzen's consistency proof within infinitary proof theory. In *Computational Logic and Proof Theory, KGC '97* (ed. by G. Gottlob et al.), Lecture Notes in Comput. Sci. 1289, Springer-Verlag, Berlin 1997, 4–17.

[10] Cantor, G., Beiträge zur Begründung der transfiniten Mengenlehre II. *Math. Ann.* **49** (1897), 207–246.

[11] Carlson, T., Elementary patterns of resemblance. *Ann. Pure Appl. Logic* **108** (2001), 19–77.

[12] Drake, F., *Set Theory: An introduction to large cardinals*. Stud. Logic Found. Math. 76, North Holland, Amsterdam 1974.

[13] Feferman, S., Systems of predicative analysis. *J. Symbolic Logic* **29** (1964), 1–30.

[14] Feferman, S., Proof theory: a personal report. In *Proof Theory* (ed. by G. Takeuti), Stud. Logic Found. Math. 81, 2nd edition, North-Holland, Amsterdam 1987, 445–485.

[15] Feferman, S., Remarks for "The Trends in Logic". In *Logic Colloquium '88*, North-Holland, Amsterdam 1989, 361–363.

[16] Friedman, H., Robertson, N., Seymour, P., The metamathematics of the graph minor theorem. *Contemp. Math.* **65** (1987), 229–261.

[17] Hardy, G. H. A theorem concerning the infinite cardinal numbers. *Quart. J. Math.* **35** (1904), 87–94.

[18] Hilbert, D., Die Grundlegung der elementaren Zahlentheorie. *Math. Ann.* **104** (1931), 485–494.

[19] Hilbert, D., and Bernays, P., *Grundlagen der Mathematik II*. Grundlehren Math. Wiss. 50, Springer-Verlag, Berlin 1939.

[20] Jäger, G., Zur Beweistheorie der Kripke–Platek Mengenlehre über den natürlichen Zahlen. *Arch. Math. Logik Grundlag.* **22** (1982), 121–139.

[21] Jäger, G., and Pohlers, W., Eine beweistheoretische Untersuchung von Δ^1_2-**CA** + **BI** und verwandter Systeme. *Bayer. Akad. Wiss. Math.-Natur. Kl. Sitzungsber.* **1982** (1983), 1–28.

[22] Kanamori, A., Magidor, M., The evolution of large cardinal axioms in set theory. In *Higher Set Theory* (ed. by G. H. Müller, D. S. Scott), Lecture Notes in Math. 669, Springer-Verlag, Berlin 1978, 99–275.

[23] Pohlers, W., Cut elimination for impredicative infinitary systems, part II: Ordinal analysis for iterated inductive definitions. *Arch. Math. Logik Grundlag.* **22** (1982), 113–129.

[24] Pohlers, W., Proof theory and ordinal analysis. *Arch. Math. Logic* **30** (1991), 311–376.

[25] Rathjen, M., Ordinal notations based on a weakly Mahlo cardinal. *Arch. Math. Logic* **29** (1990), 249–263.

[26] Rathjen, M., Proof-Theoretic Analysis of KPM. *Arch. Math. Logic* **30** (1991), 377–403.

[27] Rathjen, M., How to develop proof–theoretic ordinal functions on the basis of admissible sets. *Math. Logic Quart.* **39** (1993), 47–54.

[28] Rathjen, M., Collapsing functions based on recursively large ordinals: A well–ordering proof for KPM. *Arch. Math. Logic* **33** (1994), 35–55.

[29] Rathjen, M., Proof theory of reflection. *Ann. Pure Appl. Logic* **68** (1994), 181–224.

[30] Rathjen, M., Recent advances in ordinal analysis: Π^1_2-CA and related systems. *Bull. Symbolic Logic* **1** (1995), 468–485.

[31] Rathjen, M., The realm of ordinal analysis. In *Sets and Proofs* (ed. by S. B. Cooper and J. K. Truss), London Math. Soc. Lecture Note Ser. 258, Cambridge University Press, Cambridge 1999, 219–279.

[32] Rathjen, M., An ordinal analysis of stability. *Arch. Math. Logic* **44** (2005), 1–62.

[33] Rathjen, M., An ordinal analysis of parameter-free Π^1_2 comprehension. *Arch. Math. Logic* **44** (2005), 263–362.

[34] Richter, W. and Aczel, P., Inductive definitions and reflecting properties of admissible ordinals. In *Generalized Recursion Theory* (ed. by J. E. Fenstad and P. G. Hinman), Stud. Logic Found. Math. 79, North Holland, Amsterdam 1973, 301–381.

[35] Schütte, K., Beweistheoretische Erfassung der unendlichen Induktion in der Zahlentheorie. *Math. Ann.* **122** (1951), 369–389.

[36] Schütte, K., *Beweistheorie*. Grundlehren Math. Wiss. 103, Springer-Verlag, Berlin 1960.

[37] Schütte, K., Eine Grenze für die Beweisbarkeit der transfiniten Induktion in der verzweigten Typenlogik. *Arch. Math. Logik Grundlagenforsch.* **67** (1964), 45–60.

[38] Schütte, K., Predicative well-orderings. In *Formal systems and recursive functions* (ed. by J. N. Crossley and M. A. E. Dummett), North-Holland, Amsterdam 1965, 176–184.

[39] Simpson, S., Nichtbeweisbarkeit von gewissen kombinatorischen Eigenschaften endlicher Bäume. *Arch. Math. Logik Grundlag.* **25** (1985), 45–65.

[40] Simpson, S., *Subsystems of second order arithmetic*. Perspect. Math. Logic, Springer-Verlag, Berlin 1999.

[41] Takeuti, G., Consistency proofs of subsystems of classical analysis. *Ann. of Math.* (2) **86** (1967), 299–348.

[42] Takeuti, G., Proof theory and set theory. *Synthese* **62** (1985), 255–263.

[43] Takeuti, G., M. Yasugi, M., The ordinals of the systems of second order arithmetic with the provably Δ^1_2-comprehension and the Δ^1_2-comprehension axiom respectively. *Japan J. Math.* **41** (1973), 1–67.

[44] Veblen, O., Continuous increasing functions of finite and transfinite ordinals. *Trans. Amer. Math. Soc.* **9** (1908), 280–292.

Department of Mathematics, The Ohio State University, Columbus, Ohio 43210, U.S.A.
and
Department of Pure Mathematics, University of Leeds, Leeds LS2 9JT, England
E-mail: rathjen@math.ohio-state.edu

Analytic difference rings

Thomas Scanlon*

Abstract. Generalizing and synthesizing earlier work on the model theory of valued difference fields and on the model theory of valued fields with analytic structure, we prove Ax–Kochen–Eršov style relative completeness and relative quantifier elimination theorems for a theory of valuation rings with analytic and difference structure. Specializing our results to the case of $W[\mathbb{F}_p^{\mathrm{alg}}]$, the ring of Witt vectors of the algebraic closure of the field with p elements, given together with the relative Frobenius and the Tate algebras as analytic structure, we develop a model theoretic account of Buium's p-differential functions. In so doing, we derive a uniform p-adic version of the Manin–Mumford conjecture.

Mathematics Subject Classification (2000). Primary 03C10, 03C60, 12J10; Secondary 11D45, 11G10, 12H10, 13K05.

Keywords. Ax–Kochen–Eršov principle, difference ring, p-differential function, Witt vectors, abelian varieties, Manin–Mumford conjecture.

1. Introduction

If (K, v) is a complete valued field and $f(x_1, \ldots, x_n) = \sum a_\alpha x^\alpha \in K[[x_1, \ldots, x_n]]$ is a formal power series over K for which $v(a_\alpha) \to \infty$ as $|\alpha| \to \infty$, then f defines a function $\mathcal{O}_K^n \to K$. Considering a formal first-order language rich enough to express the field structure, the binary relation $v(x) \leq v(y)$ and the functions coming from such convergent power series, one has a natural logical setting for studying nonarchimedian analysis. If one includes in addition a unary function symbol σ to denote a field automorphism which respects the valuation in the sense that $v(x) = v(\sigma(x))$ universally and respects the analytic structure in the sense that $\sigma(f(x)) = f^\sigma(\sigma(x))$ where f^σ denotes the effect of applying σ to the coefficients of f, then one has a strong enough language to study analytic difference rings, the central object of consideration in this paper.

While we have several motivations to study these structures, two stand out most prominently. First, following the seminal work of Ax and Kochen [1], [2], [3] and Eršov on the model theory of valued fields, a great many results showing that valued fields considered in ever more complicated languages have very elegant theories have been proven. With this work we amalgamate two different strands of the model theory of enriched valued fields. Namely, we show that the theories of valued fields with analytic structure and of valued difference fields may be unified. Further unification is

*Partially supported by an NSF CAREER award.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

certainly possible. Routine modifications of the proofs presented here should suffice to combine analytic and differential structure, or more generally D-structure, while other extensions will require the development of genuinely new methods. Secondly, we wish to give a model theoretic account of Buium's theory of p-differential geometry and thereby deduce uniformities in Diophantine geometry through applications of the compactness theorem and appropriate quantifier elimination theorems.

Let us recall a little of the theory of p-differential operators. For p a prime number a p-derivation δ on a commutative ring R is a function $\delta\colon R \to R$ satisfying

- $\delta(1) = 0$,
- the functional equation $\delta(x+y) = \delta(x) + \delta(y) + \Phi_p(x, y)$ where $\Phi_p(X, Y) \in \mathbb{Z}[X, Y]$ is the integral polynomial $\frac{1}{p}(X^p + Y^p - (X+Y)^p)$, and
- the functional equation $\delta(xy) = y^p\delta(x) + x^p\delta(y) + p\delta(x)\delta(y)$.

Given a p-derivation $\delta\colon R \to R$ one can define a ring endomorphism $\sigma\colon R \to R$ by the equation $\sigma(x) := x^p + p\delta(x)$. Conversely, if p is not a zero divisor in R and $\tau\colon R \to R$ is an endomorphism lifting the Frobenius in the sense that $\tau(x) \equiv x^p \pmod{p}$ for all $x \in R$, then $\tilde{\delta}\colon R \to R$ defined by $\tau(x) = x^p + p\tilde{\delta}(x)$ is a p-derivation.

As with differential algebra, there is a p-differential geometry associated to the category of rings with p-derivations. At the naïve level, one can consider sets defined by the vanishing of p-differential polynomials, expressions of the form $P(\boldsymbol{x}, \ldots, \delta^n(\boldsymbol{x}))$ where P is a polynomial, as the basic affine sets. In the case that the underlying rings are domains, this p-differential geometry is essentially the same as the corresponding difference algebraic geometry coming from difference equations involving σ and there is already a well developed model theoretic approach to this subject [7], [8]. However, a richer geometry more in line with that of Kolchin's differential algebraic geometry may be obtained by p-adically completing the rings of p-differential polynomials. Indeed, Buium notes that to globalize p-differential geometry one must consider these p-adically complete rings of operators. The fundamental functions in this theory, the p-differential functions, locally have the form $F(\boldsymbol{x}, \ldots, \delta^n\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, \ldots, x_m)$ and F is given by p-adically convergent power series in $(n+1)m$ variables. Buium shows that many arithmetically interesting functions on the $R := W[\mathbb{F}_p^{\mathrm{alg}}]$-rational points of schemes over R may be expressed locally as p-differential functions where one takes $\delta := \frac{1}{p}(x^p - \sigma(x))$ with $\sigma\colon R \to R$ the Witt–Frobenius, the unique lifting of the Frobenius automorphism to an automorphism of the Witt vectors.

One sees from the above local description of p-differential functions, that every p-differential function over R may be expressed as a term in the language with function symbols for p-adically convergent power series over R, the Witt–Frobenius, and the restricted division function $D_p\colon R \to R$ defined by $D_p(x) := \frac{x}{p}$ if $x \in pR$ and $D_p(x) := 0$ otherwise. Conversely, if one were to regard all p-differential functions as definable, then all of the above basic functions would be definable as

well. Consequently, the logic of Buium's p-differential functions is that of the first-order structure of the Witt vectors of the algebraic closure of the field of p elements with a function symbol for the Witt–Frobenius and for all p-adic analytic functions.

Even though the goal of understanding p-differential functions guides our work, we must consider structures of a more abstract nature in order to prove our results sufficiently uniformly in order to derive any useful information about p-differential geometry. We achieve these results by axiomatizing the notion of an analytic difference structure on a valued field and then proving relative completeness and relative quantifier elimination theorems for analytic difference rings in the style of the Ax–Kochen–Eršov theorems for pure valued fields.

The essential tool in our analysis is a uniform version of the Weierstraß division theorem. Fortunately for us, this theorem is already known in the case of most interest to us [20]. Using the uniform Weierstraß division theorem we are able to assign an order-degree to an analytic difference equation with respect to which we may carry out inductive proofs.

The present author previously considered the ring $W[\mathbb{F}_p^{\mathrm{alg}}]$ simply as a difference ring in [17], [5] where a simple axiomatization was presented and a quantifier simplification theorem was proven. However, since difference polynomials are intrinsically finitistic objects, we were able to consider more complicated degree relations and worked with a version of Hensel's lemma unavailable in the analytic difference context. The restrictions imposed by considering simultaneously analytic and difference structure have forced us to employ an ostensibly weaker form of Hensel's lemma which miraculously suffices.

This paper is organized as follows. In Section 2 we introduce our basic axioms for analytic difference rings and establish some of the fundamental results about these structures. In Section 3 we state and prove our Ax–Kochen–Eršov theorems for analytically difference henselian rings. In Section 4 we recall the theory of p-differential functions in detail and apply our results of Section 3 to prove a uniform version of the Manin–Mumford conjecture.

2. Foundations of analytic and difference structure

We begin this section by recalling that a *difference ring* (R, σ) is a commutative (unital) ring R given together with a distinguished ring endomorphism $\sigma : R \to R$. While we shall usually consider rings for which σ is an automorphism, we do not insist upon this condition in our definition of the term difference ring. The model theory of difference fields, namely fields given together with a distinguished endomorphism, and, hence, also of difference domains, has been described by Chatzidakis and Hrushovski [7] and in all characteristics by Chatzidakis, Hrushovski, and Peterzil [8].

For us, a valued difference field is a valued field (K, v) given together with a distinguished automorphism $\sigma : K \to K$ which respects the valuation in the sense

that the equality $v(\sigma(x)) = v(x)$ holds universally. The model theory of valued difference fields has been developed by Bélair, Macintyre and Scanlon [17], [5].

As mentioned in the introduction, an analytic difference ring is simply the ring of integers of a valued difference field given together with analytic functions for which the distinguished automorphism respects the analytic structure. For a *fixed* complete valuation ring it is easy enough to say what one means by analytic structure. However, if one wishes to express the axioms for the theory of such a ring in a first-order language, it is necessary to formulate "analytic structure" more abstractly. Moreover, even if one is only interested in complete rings, to compare the theories of these rings as analytic structures one requires a uniform language.

We adapt van den Dries' treatment of analytic Ax–Kochen–Eršov theorems [19] and its refinements by van den Dries, Haskell, Macpherson, Lipshitz and Robinson [20], [15] to the valued difference field setting. While we could restrict our attention to such rings of analytic functions as $\mathbb{Z}[[t]]\langle X_1, \ldots, X_n\rangle$ or $W[\mathbb{F}_p^{\mathrm{alg}}]\langle X_1, \ldots, X_n\rangle$ without sacrificing the examples of greatest interest, we work with potentially more general rings in order to separate the work on the model theory of analytic functions from difference algebra.

Definition 2.1. A *pre-notion of analyticity*, $\mathcal{A}$, is given by the data of a commutative ring R and a doubly-indexed sequence of subrings $\mathcal{A}_{m,n} \subseteq R[\boldsymbol{X}][[\boldsymbol{Y}]]$ of the ring of formal power series in the n variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ over the polynomial ring in the m variables $\boldsymbol{X} = (X_1, \ldots, X_m)$ over R for which

1. $\mathcal{A}_{0,0} = R$,
2. if $m \leq m'$ and $n \leq n'$, then $\mathcal{A}_{m,n}$ is a subring of $\mathcal{A}_{m',n'}$ via the natural inclusion, and
3. $\mathcal{A}$ is closed under compositions as far as this makes sense.

Definition 2.2. Given a pre-notion of analyticity $\mathcal{A}$, an $\mathcal{A}$-analytic structure on a valuation ring $\mathcal{O}$ with maximal ideal $\mathfrak{m}$ is given by a sequence of homomorphisms $I_{m,n}\colon \mathcal{A}_{m,n} \to \mathrm{Functions}(\mathcal{O}^m \times \mathfrak{m}^n, \mathcal{O})$ which respect the compositional identities in $\mathcal{A}$, the identities coming from the inclusions $\mathcal{A}_{m,n} \hookrightarrow \mathcal{A}_{m',n'}$, and send the variables X_i and Y_i to the obvious projection maps.

Remark 2.3. If R itself is a complete valuation ring and $\mathcal{A}_{m,n} = R[\boldsymbol{X}][[\boldsymbol{Y}]]$, then the usual interpretation of the elements of $\mathcal{A}_{m,n}$ gives R an $\mathcal{A}$-analytic structure.

Remark 2.4. In the definition of $\mathcal{A}$-analytic structure, it is not really necessary that $\mathcal{O}$ be a valuation ring and $\mathfrak{m}$ its maximal ideal. However, this is the only case we consider in our applications.

Remark 2.5. Given a pre-notion of analyticity $\mathcal{A}$ and $\mathcal{L}$ a first-order language for valued fields containing (at least) a sort symbol $\mathcal{O}$ for the valuation ring and a sort symbol $\mathfrak{m}$ for the maximal ideal of the valuation ring we may naturally expand $\mathcal{L}$ to $\mathcal{L}(\mathcal{A})$ by new function symbols where for each $f \in \mathcal{A}_{m,n}$ we have a function

symbol, also denoted f, of domain sort $\mathcal{O}^m \times \mathfrak{m}^n$ and range sort $\mathcal{O}$. The condition that a particular interpretation of $\mathcal{A}$ on a valuation ring defines an $\mathcal{A}$-analytic structure may be expressed as a first-order theory in $\mathcal{L}(\mathcal{A})$.

Before we can give our conditions on when a pre-notion of analyticity is actually a notion of analyticity, we must recall some of the basic formalism of quotient operators on valuation rings and leading term structures. In what follows, we use the symbol $\mathcal{Q}$ for our quotient operators even though "D" is more common in the literature.

Definition 2.6. Let (K, v) be a valued field with valuation ring $\mathcal{O} := \mathcal{O}_{K,v}$ having the maximal ideal $\mathfrak{m} := \mathfrak{m}_{K,v}$. We define two operators $\mathcal{Q}_0$ and $\mathcal{Q}_1$ on $\mathcal{O}^2$ by $\mathcal{Q}_0(x, y) := \frac{x}{y}$ if $v(x) \geq v(y) \neq \infty$ and $\mathcal{Q}_0(x, y) = 0$ otherwise while $\mathcal{Q}_1(x, y) := \frac{x}{y}$ if $v(x) > v(y)$ and is zero otherwise.

Remark 2.7. As shown in the work of Denef and van den Dries [9] and Lipshitz and Robinson [15], for example, quantifier elimination for certain valuation rings considered with analytic structure may be obtained in languages possessing $\mathcal{Q}_0$ and $\mathcal{Q}_1$ as primitives, but not without these operators.

Definition 2.8. Given a pre-notion of analyticity $\mathcal{A}$ and a first-order language of valuation rings $\mathcal{L}$ as in Remark 2.5, the language $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$ is the expansion of $\mathcal{L}(\mathcal{A})$ by the function symbol $\mathcal{Q}_0$ and $\mathcal{Q}_1$ of domain sort $\mathcal{O}^2$ and range sorts $\mathcal{O}$ and $\mathfrak{m}$, respectively. Given a valuation ring with $\mathcal{A}$-analytic structure there is a natural expansion of the structure to an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$-structure.

We recall now the formalism of leading terms and angular components.

Definition 2.9. Let (K, v) be a valued field and $t \in \mathcal{O} = \mathcal{O}_{K,v}$ be a fixed nonzero element of the ring integers of K. For each natural number n, we define the n^{th} leading terms of K relative to K to be the multiplicative monoid $\ell_{n,t}(K) := K/(1 + t^n\mathfrak{m})$. We write $\ell_{n,t}(K)^* := \ell_{n,t}(K) \smallsetminus \{0\}$. We write $r_{n,t}(K) := \mathcal{O}/t^n\mathfrak{m}$. If t is understood we write simply $\ell_n(K)$ for $\ell_{n,t}(K)$ and $r_n(K)$ for $r_{n,t}(K)$ We write $\ell_n \colon K \to \ell_n(K)$ for the natural quotient map and $\pi_n \colon \mathcal{O} \to r_n(K)$ for the reduction map.

Remark 2.10. While $\ell_n(K)^*$ is naturally a group, it carries additional structure. For instance, the valuation map $v \colon K \to \Gamma_K \cup \{\infty\}$ descends to a map on $\ell_n(K)$ which we continue to denote by v. More importantly, addition leaves a trace on $\ell_n(K)$ in the form of a ternary predicate $\tilde{+}_n := \{(x, y, z) \in \ell_n(K)^3 \mid \exists \tilde{x}, \tilde{y}, \tilde{z} \in K\ \tilde{x} + \tilde{y} = \tilde{z}, \ell_n(\tilde{x}) = x, \ell_n(\tilde{y}) = y, \text{ and } \ell_n(\tilde{z}) = z\}$. In the sequel we shall require that $\ell_n(K)$ remember more structure from K. In particular, we insist that the leading terms remember analytic identities. That is, for each $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$-term $f(x_1, \ldots, x_m)$ the image of $\{(x_1, \ldots, x_m, y) \in K^{m+1} \mid f(\boldsymbol{x}) = y\}$ under ℓ_n is to be described by an m-ary predicate on ℓ_n.

Remark 2.11. The image of $\mathcal{O}^\times$ in $\ell_n(K)$ may be identified with $r_n(K)^\times$ and the valuation exact sequence $1 \longrightarrow \mathcal{O}^\times \longrightarrow K^\times \xrightarrow{v} \Gamma_K \longrightarrow 0$ descends to $1 \longrightarrow r_n(K)^\times \longrightarrow \ell_n(K)^* \xrightarrow{v} \Gamma_K \longrightarrow 0$.

Remark 2.12. If $t \in \mathcal{O}^\times$ is a unit, then the leading term structures $\ell_{n,t}(K)$ are all identical.

Remark 2.13. In our intended applications we take $t = p$, the residue characteristic, or $t = 1$ when the residue characteristic is zero. In fact, we shall impose this requirement with our axioms.

Remark 2.14. One can consider leading term structures relative to other ideals in $\mathcal{O}_K$ and we shall use $\ell_\infty(K) := K/(1+t^\infty\mathfrak{m})$ where $t^\infty\mathfrak{m} := \{x \in \mathcal{O} \mid (\forall n \in \mathbb{Z}_+)v(x) > v(t^n)\}$. One cannot access $\ell_\infty(K)$ directly in first-order logic, but when the field K is $\aleph_1$-compact, $\ell_\infty(K) = \varprojlim \ell_n(K)$ so that it may be approached from first-order data. Note that the ring $\mathcal{O}(K)\left[\frac{1}{t}\right]$ is a valuation ring whose residue field is $r_\infty(K)\left[\frac{1}{\pi_\infty(t)}\right]$. We refer to the corresponding coarsened valuation as v_∞.

Remark 2.15. In the work of Basarab and Kuhlmann [4], [14], leading term structures are called "additive-multiplicative congruences" or "amc structures."

Leading term structures already live definably in valued fields, but the way in which they nontrivially combine the value group and certain residue rings can complicate their analysis. By working with angular component functions one can treat these parts separately.

Definition 2.16. An *angular component function of level n* is a section $\mathrm{ac}_n : \ell_n(K)^* \to r_n(K)^\times$ of the valuation sequence. A *system of angular component functions* is a sequence $\{\mathrm{ac}_n\}_{n=0}^\infty$ where ac_n is an angular component function of level n and these functions commute with the obvious quotient maps between the leading term and residue sorts.

Remark 2.17. As with the leading terms, we shall require that the angular component functions preserve more than just the multiplicative structure.

Remark 2.18. While angular components need not exist in general, they do if (K, v) is sufficiently saturated. Thus, possibly at the cost of replacing (K, v) with an elementarily equivalent structure, we may assume that we have angular component functions.

Let us now fix once and for all a background language $\mathcal{L}$ and theory of valued fields, T_{VF}. We take $\mathcal{L}$ to be a many sorted language having sort symbols VF for the valued field itself, $\mathcal{O}$ for the valuation ring, $\mathfrak{m}$ for the maximal ideal of the valuation ring, Γ for the value group, r_n for the residue rings of Definition 2.9 and $r_n^\times$ for the units in the residue ring, and ℓ_n for the leading terms. The sorts are connected by the inclusion maps $\mathfrak{m} \hookrightarrow \mathcal{O} \hookrightarrow \mathrm{VF}$, $r_n^\times \hookrightarrow r_n$ and $r_n^\times \hookrightarrow \ell_n$, the valuation maps $v\colon \mathrm{VF} \to \Gamma$ and $v\colon \ell_n \to \Gamma$, the reduction maps $\pi_n\colon \mathcal{O} \to r_n$ and $\pi_{m,n}\colon r_m \to r_n$, and the leading term maps $\ell_n\colon \mathrm{VF} \to \ell_n$ and $\ell_{m,n}\colon \ell_m \to \ell_n$. The sorts VF, $\mathcal{O}$, and r_n come equipped with a copy of the language of rings while Γ is presented in the language of ordered abelian groups and the ℓ_n sorts each have a binary multiplication

operation and a ternary predicate for addition as described above. If we wish to include angular component functions, then expand the language to $\mathcal{L}(\{\mathrm{ac}_n\})$.

We axiomatize the theory of valued fields, T_{VF}, in $\mathcal{L}$ with the usual axioms asserting that if $M \models T$, then $\mathrm{VF}(M)$ is a field and that $v\colon \mathrm{VF}(M) \to \Gamma(M)$ is a valuation, and that all of the other sorts are interpreted as expected. That is, the inclusion maps $\mathfrak{m}(M) \hookrightarrow \mathcal{O}(M) \hookrightarrow \mathrm{VF}(M)$ are really inclusions and identify their images with the elements of positive valuation and of nonnegative valuation, respectively, the valuation maps are surjective, and the residue ring sorts and leading term sorts really give the residue rings and leading terms, *et cetera*. The one nontrivial point here is that we require $\ell_n(M)$ to be $\mathrm{VF}(M)/(1+\mathfrak{m}(M))$ ($r_n(M)$ to be $\mathcal{O}(M)/\mathfrak{m}(M)$, respectively) if the residue characteristic is zero and to be $\mathrm{VF}(M)/(1+p^n\mathfrak{m}(M))$ ($\mathcal{O}(M)/p^n\mathfrak{m}(M)$, respectively) when the residue characteristic is $p > 0$. This condition may be expressed by a set of first-order sentences. Of course, if we work in $\mathcal{L}(\{\mathrm{ac}_n\})$, then our theory $T_{\mathrm{VF}}(\mathrm{ac})$ expresses that the angular component function symbols are interpreted as angular components.

When we expand to $\mathcal{L}^{\mathcal{Q}}$ our theory $T_{\mathrm{VF}}^{\mathcal{Q}}$ includes axioms expressing the definitions of $\mathcal{Q}_0$ and $\mathcal{Q}_1$. Given a pre-notion of analyticity $\mathcal{A}$, we require of the expanded language $\mathcal{L}(\mathcal{A})$ not only that there be function symbols for the elements on $\mathcal{A}$ but that there be predicates on the leading term sorts corresponding to these functions. Given any $\mathcal{L}$-theory $T \supseteq T_{\mathrm{VF}}$ of valued fields, the theory $T(\mathcal{A})$ is obtained from T by adjoining the axioms expressing that the valuation ring has $\mathcal{A}$-analytic structure and that the new predicates on the leading terms are interpreted correctly.

Remark 2.19. For the main theorems of this paper we require that the valued fields under consideration have characteristic zero.

We need to say a little about affinoids before finishing the definition of a notion of analyticity. If $M \models T_{\mathrm{VF}}(\mathcal{A})$ is a valuation ring with $\mathcal{A}$-analytic structure and (K', v') is an algebraic extension of $\mathrm{VF}(M)$ with an extension of the valuation v, then there is a unique way to extend the $\mathcal{A}$-analytic structure to K'. Indeed, it is enough to see this in the case that K' is a finite extension of $\mathrm{VF}(M)$. Fixing a basis for $\mathcal{O}(K')$ over $\mathcal{O}(M)$, one can identify $\mathcal{O}(K')$ with $\mathcal{O}(M)^{[K:\mathrm{VF}(M)]}$. In so doing, one can expand the action of the $\mathcal{A}$-analytic functions in terms of this basis as well. In particular, if $K' = \mathrm{VF}(M)^{\mathrm{alg}}$ is the algebraic closure of $\mathrm{VF}(M)$, then $(K', v) \models T_{\mathrm{VF}}(\mathcal{A})$.

Definition 2.20. Let $M \models T_{\mathrm{VF}}(\mathcal{A})$ be a valuation ring with $\mathcal{A}$-analytic structure and fix an extension v' of v to $K' := \mathrm{VF}(M)^{\mathrm{alg}}$. A S subset of $\mathcal{O}(K')$ is said to be an *affinoid over* M if there are $\gamma_1, \ldots, \gamma_n \in \Gamma(M)$ and $a_1, \ldots, a_n \in \mathcal{O}(M)$ with $\gamma_1 > \gamma_i$ for $i \neq 1$ and $S = \{z \in \mathcal{O}(K') \mid v(z-a_1) \geq \gamma_1 \ \wedge \ \bigwedge_{i=2}^{n} v(z-a_i) \leq \gamma_i\}$. An affinoid set in M is the intersection of an affinoid set over M with $\mathcal{O}(M)$.

With the background on valued fields in place we are now ready to describe when a pre-notion of analyticity is actually a notion of analyticity.

Definition 2.21. Fix some theory $T \supseteq T_{\mathrm{VF}}$ of valued fields. A *notion of analyticity* (*relative to* T) is a pre-notion of analyticity, $\mathcal{A}$, for which Weierstraß division holds uniformly in the following sense. If $M \models T(\mathcal{A})$ and $t(x)$ is an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_M$-term in the single $\mathcal{O}$-variable x, then there are finitely many affinoid subsets $F_1, \ldots, F_n$ of $\mathcal{O}(M)$ for which $\mathcal{O}(M) = \bigcup F_i$ and for each i there is a rational function $R_i(X)$ over $\mathcal{O}(M)$ having no poles in F_i and $\mathcal{L}(\mathcal{A})_M$-terms $E_i(x)$ and $E_i^{-1}(x)$ for which $E_i(x)E_i^{-1}(x) \equiv 1$ on F_i and $t(x) = E_i(x)R_i(x)$ at all but finitely many points of F_i.

Remark 2.22. That the rings of convergent power series over complete DVRs give a notion of analyticity is proven by van den Dries, Haskell and Macpherson in [20]. (Combine Proposition 4.1 with Corollary 3.4 noting that Proposition 4.1 is still general even though it is in Section 4 where the authors claim to specialize to the case of the p-adics.)

Remark 2.23. In our applications, we restrict attention to valued fields of characteristic zero. Thus, the theory T in Definition 2.21 will be T_{VF} together with the set of sentences asserting that the valued field itself has characteristic zero.

Remark 2.24. The condition of uniform Weierstraß division may be expressed more syntactically in that the parameters for the term $t(x)$ may be given as a tuple of variables $\boldsymbol{y}$ and then the affinoids, the rational functions, and the units $E(x)$ vary uniformly with $\boldsymbol{y}$.

If R is any ring and $\sigma : R \to R$ is an automorphism, then σ extends to an automorphism $\sigma : R[\boldsymbol{X}][[\boldsymbol{Y}]] \to R[\boldsymbol{X}][[\boldsymbol{Y}]]$ of the power series ring over the polynomial ring over R. For $f \in R[\boldsymbol{X}][[\boldsymbol{Y}]]$ we write the f^{σ} for the result of applying σ to f.

Definition 2.25. A *notion of difference analyticity* (relative to T as in Definition 2.21), $(\mathcal{A}, \sigma)$, is given by a notion of analyticity $\mathcal{A}$ and an automorphism $\sigma : \mathcal{A}_{0,0} \to \mathcal{A}_{0,0}$ which induces an automorphism on each $\mathcal{A}_{m,n}$.

Definition 2.26. Given a notion of difference analyticity $(\mathcal{A}, \sigma)$ (relative to T), an $\mathcal{A}$-analytic difference ring is a model $M \models T(\mathcal{A})$ given together with a distinguished automorphism $\sigma : M \to M$ which preserves the valuation in the sense that $v(\sigma(x)) = v(x)$ universally and respects the $\mathcal{A}$-analytic structure in the sense that $\sigma(f(\boldsymbol{x})) = f^{\sigma}(\sigma(\boldsymbol{x}))$ for any $\mathcal{A}$-function f.

The condition of being an $\mathcal{A}$-analytic difference ring is clearly axiomatizable in $\mathcal{L}(\mathcal{A}, \sigma)$, the expansion of the language of valuation rings with $\mathcal{A}$-analytic structure by a symbol for an automorphism.

As in the study of difference algebra, terms in $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)$ may be expressed using terms from $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$ applied to prolongations, sequences of the form $\boldsymbol{\sigma}(\boldsymbol{x}) = (\boldsymbol{x}, \sigma(\boldsymbol{x}), \ldots, \sigma^n(\boldsymbol{x}))$. That is, if $t = t(\boldsymbol{x}) = t(x_1, \ldots, x_m)$ is an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)$ term, then we can find an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$ term $\tilde{t} = \tilde{t}(x_{0,1}, \ldots, x_{0,m}; \ldots; x_{n,1}, \ldots, x_{n,m})$ so that relative to the theory of $\mathcal{A}$-analytic difference rings we have $t(\boldsymbol{x}) = \tilde{t}(\boldsymbol{\sigma}(\boldsymbol{x}))$. We define the *order* of t to be the least m for which such a $\tilde{t}$ exists. It should be noted that

the order of t when computed in a fixed $\mathcal{A}$-analytic difference ring may be different from the order when computed relative to the theory of $\mathcal{A}$-analytic difference rings. In our applications, when we speak of *order* we mean *order relative to a given structure*.

Fix an $\mathcal{A}$-analytic difference ring M and $A \subseteq M$ a substructure for which $\mathcal{O}(A)$ generates A. If $a \in \mathcal{O}(M)$ and for some $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)_A$ term $t(x)$ over A we have $t(a) = 0$ but $t(x) \not\equiv 0$ in a neighborhood of a, then we can find such a term of minimal possible order, n, and define the *order of a over A*, $\operatorname{ord}(a/A)$, to be that minimal order. By the uniform Weierstraß division theorem, we may write $t(x)$ as $E(\sigma^n(x))R(\sigma^n(x))$ where R is a rational function over the $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$-structure A' generated by A and $a, \ldots, \sigma^{n-1}(a)$ having no poles near $\sigma^n(a)$ and E is an $\mathcal{L}(\mathcal{A})$ term over A' which is a unit near $\sigma^n(a)$. Thus, there is actually a nonzero polynomial over A' which vanishes at $\sigma^n(a)$. We define the *degree of a over A*, $\deg(a/A)$, to be the minimal degree, d, of such a polynomial. We combine these data in the pair $(\operatorname{ord}, \deg)(a/A) := (\operatorname{ord}(a/A), \deg(a/A))$ and order them lexicographically.

As with pure valued fields and some theories of valued fields with additional structure, the model companions of theories of $\mathcal{A}$-analytic difference rings are obtained by adjoining variants of Hensel's lemma (and an axiom about the existence of constants) to the theory. Unfortunately, the usual proof of Hensel's lemma breaks down when applied to $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)$ terms as the quotient operators may introduce discontinuities. However, these terms do define generically continuous functions and if one stays within the correct domain of continuity, Newton approximation techniques do work.

Proposition 2.27. *Suppose that M is an $\mathcal{A}$-analytic difference ring and let $p_0, \ldots, p_d$ be a finite sequence of $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_M$ terms with parameters from M and variables $x_0, \ldots, x_{n-1}$. Write $P(x) = \sum_{i=0}^{d} p_i(\boldsymbol{\sigma}(x))(\sigma^n(x))^i$. Abusing notation, we write $P'(x) = \sum i p_i(\boldsymbol{\sigma}(x))(\sigma^n(x))^{i-1}$. Assume that $a \in \mathcal{O}(M)$ with $v(P(a)) > 2v(P'(a))$ and that for any $\varepsilon \in \mathcal{O}(M)$ with $v(\varepsilon) \geq v(P(a)) - v(P'(a))$ one has $\ell_0(p_i(a+\varepsilon)) = \ell_0(p_i(a))$ for $i \leq d$. Then there is some $b \in \mathcal{O}(M)$ with $v(b-a) \geq v(P(a)) - v(P'(a))$ and $v(P(b)) > v(P(a))$.*

Proof. A variant of the usual proof applies. Indeed, let $\eta := \sigma^{-n}(\mathcal{Q}_0(-P(a), P'(a)))$ and set $b := a + \eta$. From our hypotheses, $v(a-b) = v(P(a)) - v(P'(a))$ and computing $P(b)$ we have

$$\begin{aligned} P(b) &= \sum p_i(a+\eta)(\sigma^n(a+\eta))^i \\ &= \sum_{i=0}^{d} p_i(a)[1+\xi_i] \sum_{j=0}^{i} \binom{i}{j} \sigma^n(a)^{i-j} \sigma^n(\eta)^j \quad \text{where } v(\xi_i) > 0 \\ &\equiv P(a) + P'(a)\sigma^n(\eta) \pmod{P(a)\mathfrak{m}(M)} \\ &\equiv 0 \pmod{P(a)\mathfrak{m}(M)} \end{aligned}$$

□

Corollary 2.28. *With the hypotheses as in Proposition 2.27, there is a maximal pseudoconvergent sequence $\{b_\alpha\}$ from $\mathcal{O}(M)$ with $b_0 = a$ and $v(P(b_\alpha))$ increasing*

with α. If in addition $\mathcal{O}(M)$ is maximally complete, then $b := \lim b_\alpha$ exists and $P(b) = 0$.

We convert the last part of this corollary into our version of henselianity for $\mathcal{A}$-analytic difference rings.

Definition 2.29. We say that the $\mathcal{A}$-analytic difference ring M is $\mathcal{A}$-analytically difference henselian if the conclusion of Corollary 2.28 holds for M. That is, given a sequence of $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_M$ terms $p_0, \ldots, p_d$ with variables $x_0, \ldots, x_{n-1}$ writing $P(x) = \sum p_i(\boldsymbol{\sigma}(x), \boldsymbol{a})\sigma^n(x)^i$ if $a \in \mathcal{O}(M)$ has the property that $v(P(a)) > 2v(P'(a))$ while for any $\varepsilon \in \mathcal{O}(M)$ with $v(\varepsilon) \geq v(P(a)) - v(P'(a))$ we have $\ell_0(p_i(\boldsymbol{\sigma}(a))) = \ell_0(p_i(\boldsymbol{\sigma}(a + \varepsilon)))$ for $i \leq d$, then there is some $b \in \mathcal{O}(M)$ with $P(b) = 0$ and $v(b - a) \geq v(P(a)) - v(P'(a))$.

Remark 2.30. It should be noted that even when there are no quotient operators, and even in the case of difference polynomials, the continuity hypothesis is nontrivial.

Visibly, the condition of being $\mathcal{A}$-analytically difference henselian is first-order expressible. In axiomatizing the theory of $\mathcal{A}$-analytically difference henselian rings, $T_{\mathcal{A}\text{-DH}}$, we impose two additional requirements beyond those of Definition 2.29. First, we insist that every model of $T_{\mathcal{A}\text{-DH}}$ be of characteristic zero. Secondly, we demand that the valued field have enough constants in the sense that for every element of the value group there is some element of the field fixed by σ and having that valuation. This last condition can be ostensibly weakened by requiring the existence of σ-fixed elements of each valuation only at the level of the leading terms. For the remainder of this paper, when we speak of an $\mathcal{A}$-analytically difference henselian ring we mean a model of $T_{\mathcal{A}\text{-DH}}$ where $(\mathcal{A}, \sigma)$ is some notion of difference analyticity.

3. AKE theorems for analytically difference henselian rings

In this section we state and prove our main relative completeness and quantifier elimination theorems for $\mathcal{A}$-analytically difference henselian rings. As with much of the earlier work on pure valued fields and on algebraic valued difference and differential fields (but, remarkably, unlike most previous work on the model theory of analytic functions on valued fields) we prove our results by employing a model theoretic test for completeness and quantifier elimination involving extensions of partial isomorphisms.

Simply put, our theorem is that for a fixed notion of difference analyticity, $(\mathcal{A}, \sigma)$, the theory $T_{\mathcal{A}\text{-DH}}$ of $\mathcal{A}$-analytically difference henselian rings is complete and eliminates quantifiers relative to the leading term sorts, and even, resplendently so. As we expect the meaning here of relativity and resplendence may require some explanation, we describe these terms now before announcing our theorem in its official formulation.

Given a many sorted language $\mathcal{L}$ and a nonempty set Σ of $\mathcal{L}$-sort symbols, the restriction of $\mathcal{L}$ to Σ, $(\mathcal{L} \restriction \Sigma)$, is the language having sort symbols Σ and as basic function, relation, and constant symbols exactly those from $\mathcal{L}$ which refer only to sorts in Σ. That is, a function symbol f of $\mathcal{L}$ is a function symbol of $(\mathcal{L} \restriction \Sigma)$ just in case its domain sort is a sequence of sorts from Σ and its range sort belongs to Σ while a relation symbol of $\mathcal{L}$ belongs to the restricted language if its field sort is a sequence of elements of Σ and an $\mathcal{L}$-constant symbol is an $(\mathcal{L} \restriction \Sigma)$-constant if its sort belongs to Σ. If M is an $\mathcal{L}$-structure, then the restriction of M to Σ is simply the $(\mathcal{L} \restriction \Sigma)$-structure $(M \restriction \Sigma)$ consisting of the M-interpretation of the sorts in Σ and the nonlogical $(\mathcal{L} \restriction \Sigma)$-symbols.

Definition 3.1. Given a many sorted language $\mathcal{L}$ and a nonempty set Σ of $\mathcal{L}$-sort symbols we say that the $\mathcal{L}$-theory T is *complete relative to* Σ if for any model $M \models T$ the theory $T \cup \mathrm{Th}_{(\mathcal{L}\restriction\Sigma)}(M \restriction \Sigma)$ is complete.

To discuss relative quantifier elimination we need to recall Morleyization. Given a language $\mathcal{L}$, the Morleyization $\mathcal{L}^{\mathrm{Mor}}$ of $\mathcal{L}$ is obtained by adjoining to $\mathcal{L}$ a new relation symbol $R_\phi(x_1, \ldots, x_n)$ for each $\mathcal{L}$-formula ϕ with the free variables $x_1, \ldots, x_n$. The $\mathcal{L}^{\mathrm{Mor}}$-theory $T_{\mathcal{L}}^{\mathrm{Mor}}$ is defined by

$$T_{\mathcal{L}}^{\mathrm{Mor}} := \{\forall x_1 \cdots \forall x_n (R_\phi(\boldsymbol{x}) \leftrightarrow \phi(\boldsymbol{x})) \mid \phi \text{ an } \mathcal{L}\text{-formula}\}$$

On general grounds, any extension of $T_{\mathcal{L}}^{\mathrm{Mor}}$ in $\mathcal{L}^{\mathrm{Mor}}$ eliminates quantifiers.

Definition 3.2. Given a many sorted language $\mathcal{L}$ and a nonempty set Σ of $\mathcal{L}$-sort symbols we say that the $\mathcal{L}$-theory T *eliminates quantifiers relative to* Σ if the theory $T \cup T^{\mathrm{Mor}}_{(\mathcal{L}\restriction\Sigma)}$ eliminates quantifiers in $\mathcal{L} \cup (\mathcal{L} \restriction \Sigma)^{\mathrm{Mor}}$.

We mentioned that our theorems hold resplendently. We employ this enhancement of the theorem when discussing angular components. Essentially, by resplendent relative completeness (respectively, resplendent relative quantifier elimination) we mean that relative completeness (respectively, relative quantifier elimination) continues to hold even after arbitrarily enriching the sorts to which we relativize.

Definition 3.3. Let $\mathcal{L}$ be a many sorted language and Σ a nonempty set of $\mathcal{L}$-sort symbols. We say that the $\mathcal{L}$-theory T is resplendently complete relative to Σ (respectively, resplendently eliminates quantifiers relative to Σ) if for any expansion $\mathcal{L}' \supseteq (\mathcal{L} \restriction \Sigma)$ having only Σ as sort symbols and any $\mathcal{L}'$-theory T' the theory $T \cup T'$ is complete relative to Σ (respectively, eliminates quantifiers relative to Σ).

With this general nonsense on many sorted languages in place we may now state our main theorem.

Theorem 3.4. *The theory $T_{\mathcal{A}\text{-DH}}$ of $\mathcal{A}$-analytically difference henselian rings is resplendently complete relative to the leading terms sorts and resplendently eliminates quantifiers relative to the leading terms.*

Using general results on the existence of angular components, we deduce a stronger relative completeness theorem from Theorem 3.4.

Theorem 3.5. *The theory $T_{\mathcal{A}\text{-DH}}$ of $\mathcal{A}$-analytically difference henselian rings is complete relative to the value group and residue ring sorts.*

As a particular application of Theorem 3.5 we see that if $k \hookrightarrow k'$ is an extension of algebraically closed fields of characteristic p, then the corresponding extension of rings of Witt vectors, $W[k] \hookrightarrow W[k']$, is elementary in the language $\mathcal{L}(\mathcal{A}, \sigma)$ where $\mathcal{A}_{m,n} = W[k][\boldsymbol{X}][[\boldsymbol{Y}]]$ and σ is interpreted as the Witt–Frobenius. We shall expand on this observation and exploit it in Section 4.

As is our wont, we shall prove Theorem 3.4 by converting it into a statement about extending isomorphisms and then actually proving the statement on extensions by considering the cases of residue field, totally ramified, and immediate extensions separately. Some of these steps require merely routine modifications to the proofs in the algebraic setting, but others are considerably trickier.

The reader should consult Section 7 of [17] or Theorem 8.4.1 of [11] for a discussion of why the following technical theorem is equivalent to Theorem 3.4.

Theorem 3.6. *Let $\mathcal{L}'$ be an expansion of the restriction of $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)$ to the leading term sorts, $(\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma) \restriction \mathrm{LT})$, having no new sort symbols. Suppose that M_1 and M_2 are two saturated $\mathcal{A}$-analytically difference henselian rings each of the same cardinality $> (|\mathcal{L}'|^{\aleph_0})^+$ considered in the language $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma) \cup \mathcal{L}'^{\mathrm{Mor}}$. Suppose moreover that $(M_1 \restriction \mathrm{LT}) \equiv_{\mathcal{L}'} (M_2 \restriction \mathrm{LT}) \models T^{\mathrm{Mor}}_{\mathcal{L}'}$. Suppose that $A_1 \subseteq M_1$ and $A_2 \subseteq M_2$ are two small (of cardinality at most $|\mathcal{L}'|$) substructures of M_1 and M_2 for which $\mathcal{O}(A_i)$ generates A_i and that $f\colon A_1 \to A_2$ is an isomorphism of $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma) \cup (\mathcal{L}')^{\mathrm{Mor}}$-structures. If $a \in \mathcal{O}(M_1)$ is any element, then there is an extension of f to an isomorphism between the substructure of M_1 generated by A_1 and a, $A_1\langle a\rangle$, and a substructure of M_2.*

Throughout the remainder of this section we concentrate on proving Theorem 3.6, and, hence, also Theorem 3.4. In the course of this proof we shall reduce the problem to other statements with stronger hypotheses. As these restrictions are established, we shall display our new hypotheses as boxed statements.

As M_1 and M_2 are saturated of the same cardinality and $(M_1 \restriction \mathrm{LT})$ and $(M_2 \restriction \mathrm{LT})$ are elementarily equivalent, they are actually isomorphic. Since the map $f\colon A_1 \to A_2$ is an isomorphism of $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma) \cup (\mathcal{L}')^{\mathrm{Mor}}$-structures, and, each $(M_i \restriction \mathrm{LT})$ eliminates quantifiers, the restrictions of these structures to the leading terms are actually isomorphic over f. Let us fix such an isomorphism $\tilde{f}\colon (M_1 \restriction \mathrm{LT}) \to (M_2 \restriction \mathrm{LT})$ and thereby arrive at our first reduction.

$$\boxed{\begin{array}{c}\tilde{f} \cup f\colon A_1 \cup (M_1 \restriction \mathrm{LT}) \to A_2 \cup (M_2 \restriction \mathrm{LT}) \text{ is an isomorphism of} \\ \mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma) \cup (\mathcal{L}')^{\mathrm{Mor}}\text{-structures.}\end{array}}$$

We fix now $N_1 \preceq M_1$ and $N_2 \preceq M_2$ two $|\mathcal{L}'|^+$-compact elementary substructures each of cardinality less than that of M_1 for which $A_1\langle a\rangle \subseteq N_1$ and $A_2 \subseteq N_2$. Our hypotheses on the saturation of M_1 and M_2 and on their cardinalities ensure that such structures exists. In the course of our construction of an extension of f we shall initially extend inside N_1 taking values in N_2 until N_i is an immediate extension of A_i and then we extend f to a maximal immediate extension of N_1 inside M_1.

With the next lemma we show that the map f may be extended so as to add new elements to the residue ring $r_\infty(A_1)$.

Lemma 3.7. *Let $a \in r_\infty(N_1)$. Suppose that* $(\mathrm{ord}, \deg)(a/r_\infty(A_1)) = (m, d)$. *Let $\tilde{p}_0, \ldots, \tilde{p}_d$ be $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_{A_1}$ terms in the variables $x_0, \ldots, x_{n-1}$ for which the reductions under π_∞, $p_i = \pi_\infty(\tilde{p}_i)$, give well-defined functions at $(a, \sigma(a), \ldots, \sigma^{m-1}(a))$ and $P(a) = \sum p_i(\boldsymbol{\sigma}(a))(\sigma^m(a))^i = 0$. Let $\tilde{P} = \sum \tilde{p}_i(\boldsymbol{\sigma}(x))(\sigma^m(x))^i$. Then there are elements $\tilde{a} \in \mathcal{O}(N_1)$ and $\tilde{b} \in \mathcal{O}(N_2)$ for which $\tilde{P}(\tilde{a}) = 0$, $\tilde{P}^f(\tilde{b}) = 0$, $\pi_\infty(\tilde{a}) = a$, $\tilde{f}(a) = \pi_\infty(\tilde{b})$, and f extends to an isomorphism defined on the structure $A_1\langle\tilde{a}\rangle$ which has no new elements in its value group taking $\tilde{a}$ to $\tilde{b}$.*

Proof. Let $a' \in \mathcal{O}(N_1)$ be any lifting of a. By our minimality assumption, $P'(a) \neq 0$. As $P(a) = 0$ in $r_\infty(N_1)$, we see that, $2v(\tilde{P}'(a')) < v(\tilde{P}(a'))$. Moreover, because the terms $p_i(\boldsymbol{\sigma}(x))$ are well-defined at a, their leading terms do not depend on the choice of a'. Hence, as N_1 is $\mathcal{A}$-analytically difference henselian, there is some $\tilde{a}$ lifting a and satisfying $\tilde{P}(\tilde{a}) = 0$. Likewise, using $\aleph_1$-compactness of N_2 we can find $\tilde{b} \in \mathcal{O}(N_2)$ with $\tilde{P}^f(\tilde{b}) = 0$ and $\pi_\infty(\tilde{b}) = \tilde{f}(a)$.

We argue by induction on $n \leq m$ that if Q is a term of order n, then $\tilde{f}(\ell_\infty(Q(\tilde{a}))) = \ell_\infty(Q^f(\tilde{b}))$. By the uniform Weierstraß property, we may express Q near $\tilde{a}$ as $E(\sigma^n(\tilde{a}))R(\sigma^n(\tilde{a}))$ where E is given by an $\mathcal{L}(\mathcal{A})$ term over the $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$-structure generated by A_1 and $\tilde{a}, \ldots, \sigma^{n-1}(\tilde{a})$ and is a unit near $\sigma^n(\tilde{a})$ and R is a rational function over the same structure having no poles near $\sigma^n(a)$. In the case that $n = m$, we may assume that the degrees of the numerator and denominator of R are less than d. By induction, the ∞-leading terms of the parameters for E and R are under control. As the quotient operators are not applied to $\sigma^n(a)$ in E and E is a unit near $\sigma^n(a)$, its ∞-leading term is determined by that of $\sigma^n(\tilde{a})$. Write $R(\sigma^n(\tilde{a})) = S(\sigma^n(\tilde{a}))/T(\sigma^n(\tilde{a}))$ where S and T are polynomials. Write $S = c\tilde{S}$ where $v(c)$ is equal to the Gauß valuation of S. Then $\pi_\infty\tilde{S}$ gives a nonvanishing polynomial at $\sigma^n(a)$ as either $n < \mathrm{ord}(a/r_\infty(A_1))$ or $\deg \pi_\infty(\tilde{S}) < \deg(a/r_\infty(A_1))$. Thus, $\ell_\infty(\tilde{S}(\tilde{a})) = \pi_\infty(\tilde{S})(a)$. Applying the same reasoning to T, we conclude the induction and, hence, also the proof of this lemma. □

Repeatedly applying Lemma 3.7 we can extend f so that $r_\infty(A_i) = r_\infty(N_i)$. However, we delay doing this until we have achieved $\Gamma(A_i) = \Gamma(N_i)$.

With the following steps we enlarge the value group of A_1. Before actually adding new elements to the value group, we extend f so that its domain has enough constants.

Lemma 3.8. *If $c \in \mathcal{O}(A_1)$, then f extends to some $A_1\langle\varepsilon\rangle \subseteq N_1$ where $v(\varepsilon) = v(c)$, $\sigma(\varepsilon) = \varepsilon$, and $\Gamma(A_1) = \Gamma(A_1\langle\varepsilon\rangle)$.*

Proof. Take $\zeta \in \mathcal{O}(N_1)$ with $\sigma(\zeta) = \zeta$ and $v(\zeta) = v(c)$. Such an element exists by our axiom that $\mathcal{A}$-analytically difference henselian rings have enough constants. Set $\eta := \mathcal{Q}_0(\varepsilon, c)$. Then η is a nonzero solution to the linear difference equation $\sigma(X) - \mathcal{Q}(c, \sigma(c))X = 0$, even upon reduction to $r_\infty(N_1)$. Thus, there are infinitely many solutions to $\sigma(X) - \pi_\infty(\mathcal{Q}(c, \sigma(c))X = 0$ in $r_\infty(N_1)$ and by $|\mathcal{L}'|^+$-compactness, at least $|\mathcal{L}'|^+$ many solutions. In particular, there some solution a which is not algebraic over $r_\infty(A_1)$. Let $\tilde{a}$ and $\tilde{b}$ be given by Lemma 3.7 applied to $\tilde{P}(X) = \sigma(X) - \pi_\infty(\mathcal{Q}(c, \sigma(c))$. Set $\varepsilon := \tilde{a} \cdot c$. □

Iterating this construction so as to consider all the elements of the value group of A_1, we may suppose that A_1 and A_2 have enough constants.

A_1 and A_2 have enough constants

For purely ramified extensions we consider the cases of algebraic and transcendental extensions separately.

Lemma 3.9. *If $\varepsilon \in \mathcal{O}(N_1)$, $\sigma(\varepsilon) = \varepsilon$, $\varepsilon^n =: \zeta \in \mathcal{O}(A_1)$, and $mv(\varepsilon) \notin \Gamma(A_1)$ for $m < n$, then f extends to $A_1\langle\varepsilon\rangle$.*

Proof. That there is some $\eta \in N_2$ fixed by σ with $\eta^n = f(\zeta)$ and that the map extending f sending ε to η preserves the valued difference structure is already known from the algebraic case. As noted in Section 2, the $\mathcal{A}$-analytic structure extends uniquely to algebraic extensions. □

We extend now to transcendental expansions of the value group.

Lemma 3.10. *If $\varepsilon \in \mathcal{O}(N_1)$ is fixed by σ and $nv(\varepsilon) \notin \Gamma(A_1)$ for all $n \in \mathbb{Z}_+$, then there is some $\eta \in \mathcal{O}(N_2)$ also fixed by σ with $\tilde{f}(\ell_\infty(\varepsilon)) = \ell_\infty(\eta)$ for which f extends to $A_1\langle\varepsilon\rangle$ via $\varepsilon \mapsto \eta$.*

Proof. Let $\zeta \in \mathcal{O}(N_2)$ be any element with $\ell_\infty(\zeta) = \tilde{f}(\ell_\infty(\varepsilon))$ and let $P(X) = \sigma(X) - X$. Then $v(P(\zeta)) > v(\zeta) > 0 = v(P'(\zeta))$ as the leading term of ζ is a constant. Indeed, we even have $v_\infty(P(\zeta)) > v_\infty(\zeta)$ where v_∞ is the coarsened valuation. It follows that if $\xi \in \mathcal{O}(N_2)$ with $v(\xi) \geq v(P(\zeta))$, then $\ell_\infty(-(\zeta + \xi)) = \ell_\infty(-\zeta)$. Hence, our version of Hensel's lemma applies and we can find some $\eta \in \mathcal{O}(N_2)$ with $\ell_\infty(\eta) = \ell_\infty(\zeta) = \tilde{f}(\ell_\infty(\varepsilon))$ and $\sigma(\eta) = \eta$.

Since $\sigma(\varepsilon) = \varepsilon$, every element of $\mathcal{O}(A_1\langle\varepsilon\rangle)$ can be expressed as an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_{A_1}$ term applied to ε. Likewise, the same is true of η with A_1 replaced by A_2. So, it suffices to show that if $t(x)$ is an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})_{A_1}$ term, then $\tilde{f}(\ell_\infty(t(\varepsilon))) = \ell_\infty(t^f(\eta))$.

Using uniform Weierstraß division to express $t(x)$ as $E(x)R(x)$ where E is an $\mathcal{L}(\mathcal{A})_{A_1}$ term which is a unit near ε and $R(x)$ is a rational function over $\mathcal{O}(A_1)$, we see that $\ell_\infty(E(\varepsilon))$ depends just on $\ell_\infty(\varepsilon)$ and if $R(x) = \big(\sum a_i x^i\big)/\big(\sum b_j x^j\big)$, then $\ell_\infty(R(\varepsilon)) = \ell_\infty(a_{i_0})\ell_\infty(\varepsilon)^{i_0 - j_0}\ell_\infty(b_{j_0})$ where $v(a_{i_0}) + i_0 v(\varepsilon) = \min_i v(a_i) + i v(\varepsilon)$ and $v(b_{j_0}) + j_0 v(\varepsilon) = \min_j v(b_j) + j v(\varepsilon)$. □

Applying Lemmata 3.9 and 3.10 repeatedly, alternating the rôles of N_1 and N_2, we may extend f so that $\Gamma(A_i) = \Gamma(N_i)$ for $i \in \{0, 1\}$. Once this has been achieved, we may apply Lemma 3.7 repeatedly to extend f so that N_i is an immediate extension of A_i.

Let us state this result as our second reduction.

$$\boxed{N_i \text{ is an immediate extension of } A_i \text{ for } i \in \{0, 1\}}$$

Fix now $\widehat{N}_1$ a maximal immediate extension of N_1 in M_1 and a maximal immediate extension $\widehat{N}_2$ of N_2 in M_2. We shall actually extend f to $\widehat{N}_1$.

Working by induction in $\widehat{N}_1$ we may assume that a has the least possible (ord, deg) over A_1 of new elements of N_1. That is:

$$\boxed{\begin{array}{c}\text{If } b \in \mathcal{O}(N_1) \text{ and } (\text{ord}, \deg)(b/A_1) < (n, d) := (\text{ord}, \deg)(a/A_1),\\ \text{then } b \in \mathcal{O}(A_1).\end{array}}$$

Working by induction further we may assume that whenever we have a pseudo-convergent solution to a low (ord, deg) $\mathcal{A}$-analytic difference equation over A_1 in N_1, then we have an actual solution.

$$\boxed{\begin{array}{c}\text{If } Q(x) = \sum_{i=0}^{e} q_i(x, \sigma(x), \ldots, \sigma^{m-1}(x))(\sigma^m(x))^i \text{ where } (m, e) < (n, d) \text{ and}\\ \text{each } q_i \text{ is an } \mathcal{L}^{\mathcal{Q}}(\mathcal{A})_{A_1} \text{ term and } \{x_\alpha\} \text{ is a pseudosolution to } Q \text{ in the sense that}\\ v(Q(x_\alpha)) \text{ is increasing with } \alpha \text{ and Hensel's lemma applies at each } \alpha, \text{ then there}\\ \text{is some } b \in \mathcal{O}(\widehat{N}_1) \text{ with } x_\alpha \text{ pseudoconverging to } b \text{ and } Q(b) = 0.\end{array}}$$

We fix now a maximal pseudoconvergent approximation $\{x_\alpha\}$ to a from $\mathcal{O}(A_1)$ and $P(X) = \sum_{i=0}^{d} p_i(\boldsymbol{\sigma}(X))(\sigma^n(X))^i$ a minimal equation for a over A_1. We shall show the following.

1. For Q of lower complexity than that of P (that is, Q is a polynomial in $\sigma^n(X)$ of degree less than d having coefficients which are $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)_{A_1}$ terms of order less than n) we have $\ell_\infty(Q(a)) = \ell_\infty(Q(x_\alpha))$ for $\alpha \gg 0$.

2. Indeed, we shall show that $v_\infty(Q(a) - Q(x_\alpha)) > v(Q(a)) + v(a - x_\alpha)$ for $\alpha \gg 0$.

3. Possibly replacing P with a refinement, Hensel's lemma applies along $\{x_\alpha\}$.

4. There is some $b \in M_2$ for which $\{f(x_\alpha)\}$ is a pseudoconvergent approximation and $P(b) = 0$.

It follows that we may extend f by sending a to b and that our inductive stipulations on f and $\widehat{N}_1$ continue to hold.

We work by induction on $m = \text{ord}(Q)$ to prove the first of these points.

We observe first that if U is an $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma)_{A_1}$ term of order less than m, then for $\alpha \gg 0$ we have $v(a - x_\alpha) > v(\sigma^m(a) - U(a))$. Indeed, take α large enough so that the valuation inequality stated in Part 2 above holds for U. Assuming that

$v(\sigma^m(a)-U(a)) \geq v(a-x_\alpha) = v(\sigma^m(x_\alpha-a))$, then we have $v(\sigma^m(x_\alpha)-U(x_\alpha)) = v((\sigma^m(a)-U(a))+(\sigma^m(x_\alpha-a))+(U(x_\alpha)-U(a))) \geq v(a-x_\alpha)$. Hensel's lemma then applies to produce some y_α with $\sigma^m(y_\alpha) = U(y_\alpha)$ and $v(y_\alpha - x_\alpha) \geq v(a-x_\alpha)$. But then by the boxed reduction above, the sequence $\{y_\alpha\}$, and hence also $\{x_\alpha\}$ has a pseudolimit in $\widehat{N}_1$ contradicting its maximality.

The point of this observation is that any affinoid defined over the $\mathcal{L}^{\mathcal{Q}}(\mathcal{A})$-structure generated by A_1 and $a, \ldots, \sigma^{m-1}(a)$ containing $\sigma^m(a)$ also contains points of the form $\sigma^m(x_\alpha)$.

Near $\sigma^m(a)$ we can write

$$Q(x) = E(a, \ldots, \sigma^{m-1}(a); \sigma^m(x))R(a, \ldots, \sigma^{m-1}(a); \sigma^m(x))$$

where the quotient operators are not applied to $\sigma^m(x)$ in E and E is a unit near a and R is a rational function in $\sigma^m(x)$ with no poles near $\sigma^m(a)$. By induction, the parameters in E and R have the same ∞-leading terms if $(a, \ldots, \sigma^{m-1}(a))$ is replaced by $(x_\alpha, \ldots, \sigma^{m-1}(x_\alpha))$ for $\alpha \gg 0$.

Since $\ell_\infty(x_\alpha) = \ell_\infty(a)$, E is a unit, and the ∞-leading terms of the coefficients of $E(x_\alpha, \ldots, \sigma^{m-1}(x_\alpha); X)$ and $E(a, \ldots, \sigma^{m-1}(a); X)$ are the same, it follows that $\ell_\infty(E(\boldsymbol{\sigma}(a))) = \ell_\infty(E(\boldsymbol{\sigma}(x_\alpha)))$. Moreover, since the quotient operator is not applied to the last variable and $v(E(\boldsymbol{\sigma}(x_\alpha))) = 0$, the usual Taylor series expansion can be used to see that $v(E(\boldsymbol{\sigma}(x_\alpha)) - E(\boldsymbol{\sigma}(a))) = \gamma + Nv(x_\alpha - a)$ for some fixed $\gamma \in v(\mathcal{O}(A_1))$.

We can write R as $U(x_0, \ldots, x_m)/V(x_0, \ldots, x_m)$ where each of U and V is a polynomial in x_m. Let us write $U = \sum u_i(x_0, \ldots, x_{m-1})x_m^i$. By induction we know, among other things, that $\ell_\infty(u_i(\boldsymbol{\sigma}(a))) = \ell_\infty(u_i(\boldsymbol{\sigma}(x_\alpha)))$ for all i and $\alpha \gg 0$. Replacing U with $\mathcal{Q}(U, c)$ where $c \in \mathcal{O}(A_1)$ and $v(c) = \min_i v(u_i(\boldsymbol{\sigma}(a)))$ we may assume that $v(u_i(\boldsymbol{\sigma}(a))) = 0$ for some i.

Write $x_\alpha = a + y_\alpha$.

Let us expand $U(x_\alpha)$.

$$\begin{aligned} U(x_\alpha) &= \sum u_i(\boldsymbol{\sigma}(a + y_\alpha)(\sigma^m(a + y_\alpha))^i \\ &= \sum_{i,j} u_i(\boldsymbol{\sigma}(a))[1 + \xi_i]\binom{i}{j}\sigma^m(a)^{i-j}\sigma^m(y_\alpha)^j \quad \text{where } v_\infty(\xi_i) > 0 \\ &= \sum_j [1 + \zeta_j]\frac{1}{j!}U^{(j)}(a)\sigma^m(y_\alpha)^j \quad \text{where } v_\infty(\zeta_j) > 0. \end{aligned}$$

For $\alpha \gg 0$, the summands on the righthand side of the equation all have different v_∞ valuations. Thus, $\ell_\infty(U(x_\alpha)) = \ell_\infty(\frac{1}{j!}U^{(j)}(a)\sigma^m(y_\alpha)^j)$ for the j which minimizes the valuation of the expression on the right. If this j is zero, then we are done. As we have reduced to the case that $v(u_i(\boldsymbol{\sigma}(a))) = 0$ for some i, it follows that the j for which the valuation is minimized must have $v_\infty(U^{(j)}(a)) = 0$. Writing $j = k + 1$, we see that Hensel's lemma applies to $U^{(k)}(X)$ along x_α so that by our inductive hypothesis the sequence x_α pseudoconverges to a solution to $U^{(k)}(x) = 0$ contradicting its maximality. Hence, $\ell_\infty(U(a)) = \ell_\infty(U(x_\alpha))$.

Repeating this calculation with V in place of U, we finish the proof of points 1. and 2.

The above calculations apply as well to the case that $U = P$. This time since $P(a) = 0$, necessarily the minimal valuation of a summand on the right is obtained for some $j > 0$. If this j is not one and even if $v_\infty(P'(a)) \neq 0$, then as above x_α pseudoconverges to a solution of some derivative of P. Thus, Hensel's lemma applies to P along x_α and we can find the requisite solution to $P^f(X) = 0$ in $\mathcal{O}(M_2)$.

Conversely, the above calculations show that if we assumed merely that $a \in \mathcal{O}(M_1)$, then there is an immediate extension of $\widehat{N}_1$ in which x_α pseudoconverges to a solution to $P(X) = 0$. Indeed, arguing by induction on $(\text{ord}, \deg)(a/\widehat{N}_1)$ we may assume that P is also a minimal equation for a over $\widehat{N}_1$. With the above calculations we never invoked the fact that a lives in an immediate extension of A_1. Therefore, $\ell_\infty(Q(a)) = \ell_\infty(Q(x_\alpha)) \in \ell_\infty(A_1)$ for each lower complexity Q.

With these observations we conclude the proof of Theorem 3.6.

4. Model theory of p-differential geometry

In this section we apply the results from Section 3 to the theory of p-differential functions obtaining amongst other theorems a uniform version of the Manin–Mumford conjecture over $W[\mathbb{F}_p^{\text{alg}}]$.

Before discussing applications to p-differential functions we verify that the Witt vectors may indeed be regarded as $\mathcal{A}$-analytic difference henselian rings.

The reader may wish to consult Section 17 of [10] for more details on the Witt vectors. Recall that there is a functor W taking a perfect field k of characteristic $p > 0$ and returning a complete valuation ring $W[k]$ whose maximal ideal is generated by p and whose residue field is naturally isomorphic to k. From the functoriality of the Witt vector construction it follows that the Frobenius automorphism $\tau : k \to k$ induces an automorphism $W(\tau) : W[k] \to W[k]$ which reduces to τ modulo p. We refer to $W(\tau)$ as the *Witt–Frobenius*. It follows from the construction of $W(\tau)$ that it preserves the p-adic valuation on $W[k]$.

There is more than one reasonable choice for the analytic structure on $W[k]$. If we fix k, then we may wish to take $\mathcal{A}_{m,n} := W[k][\boldsymbol{X}][[\boldsymbol{Y}]]$. In this way we recover the rings of convergent power series by specializing the variables ranging over the maximal ideal. If we wish to work uniformly in p, then we may prefer to use $\mathbb{Z}[\boldsymbol{X}][[\boldsymbol{Y}]]$. In any case, the uniform Weierstraß division property follows from the main results of [20].

The most natural angular component structure on the Witt vectors is defined by taking the powers of p as the constant representatives of the value group. Henceforth, when we consider the Witt vectors with angular components we insist upon this choice. Fixing a choice of $\mathcal{A}$ as in the previous paragraph, we find now that the theory of $W[k]$ in $\mathcal{L}^{\mathcal{Q}}(\mathcal{A}, \sigma, \text{ac})$ is determined by the theory of k and admits quantifier

elimination relative to k and the value group. From Theorem 3.4 we require the theories of all of the residue rings to determine the full theory of the $\mathcal{A}$-analytic difference henselian ring. In the case of the Witt vectors, the intermediate quotients $r_n(W[k]) = W[k]/p^{n+1}W[k]$ are uniformly interpretable as the k-rational points of ring schemes over k. Thus, their theories and questions about quantifier elimination for these rings are determined by k.

Let us note two consequences of this characterization of the theory of the Witt vectors as an $\mathcal{A}$-analytic difference ring. First, if $k \hookrightarrow k'$ is an elementary extension of perfect fields, then $W[k] \hookrightarrow W[k']$ is also elementary. Secondly, the residue field is orthogonal to the value group in the sense that if $X \subseteq r_0(W[k])^n \times \Gamma(W[k])^m$ is any definable set, then X is a finite Boolean combination of sets of the form $Y \times Z$ where $Y \subseteq k^n$ is definable in k and $Z \subseteq \mathbb{Z}^m$ is definable in $(\mathbb{Z}, +, 0, <)$.

Let us turn now to a model theoretic study of p-differential geometry. As we noted in the introduction, if $\sigma\colon W[k] \to W[k]$ is the Witt–Frobenius, then the operator $\delta\colon W[k] \to W[k]$ defined by $\delta(x) := \frac{1}{p}(\sigma(x) - x^p)$ is a p-derivation and the functions of the form $f(\boldsymbol{x}) = F(\boldsymbol{x}, \delta\boldsymbol{x}, \ldots, \delta^m\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and F is given by a convergent power series in $n(m+1)$ variables are the p-differential functions on $W[k]^n$. We concentrate on one class of p-differential functions constructed by Buium, namely the p-differential characters on abelian varieties.

As an illustration of the method, we prove a uniform version of the Manin–Mumford conjecture for abelian varieties over $W[k]$. Recall that the Manin–Mumford conjecture (or Raynaud's theorem [16]) asserts that if A is an abelian variety over an algebraically closed field K of characteristic zero and $X \subseteq A$ is a closed subvariety, then the intersection of $X(K)$ with the torsion subgroup of $A(K)$ is a finite union of translates of the torsion subgroups of group subvarieties of A. For the purposes of giving this theorem a more quantitative form it can help to present it in terms of the Ueno locus of X.

Recall that the Ueno locus of X, $\mathrm{Ueno}(X)$, is the subvariety of X defined by $x \in \mathrm{Ueno}(X)(K)$ if and only if there is an abelian subvariety $B \leq A$ for which $x + B \subseteq X$. We shall have occasion to use the fact, noted in [13], that if the variety X varies in an algebraic family, then so does $\mathrm{Ueno}(X)$. The Manin–Mumford conjecture implies that there are only finitely many torsion points in $X(K)$ which do not lie in $\mathrm{Ueno}(X)(K)$. In fact, if one establishes this finiteness result for the number of torsion points lying on varieties outside their Ueno loci, then the Manin–Mumford statement follows formally.

With our terms defined we can state our uniform version of the Manin–Mumford conjecture.

Theorem 4.1. *Let k be an algebraically closed field of characteristic $p > 2$, S a variety (reduced, integral scheme of finite type) over $W[k]$ and $A \to S$ an abelian scheme over S. Let $X \subseteq A$ be a closed subscheme. Then there is a natural number N such that for any point $s \in S(W[k])$ the number of torsion points in $A_s(W[k])$ lying in $X_s(W[k])$ but outside of $\mathrm{Ueno}(X_s)(W[k])$ is bounded by N.*

Remark 4.2. The restriction to odd p is an artifact of our proof in that this is an hypothesis for the published theorem of Buium on the existence of p-differential characters.

Remark 4.3. Theorem 4.1 is similar to the main theorem of [18] but is incomparable in terms of its strength. The result in [18] is weaker in that one requires $A \to S$ to be a universal abelian variety over a moduli space and one obtains information only about fibres which are canonical lifts, but it is stronger in the sense that Zariski closure of the intersection of $X(W[k])$ with the set of torsion points on canonical lift fibres is described with greater precision than is possible under the hypotheses of Theorem 4.1.

As with some of the other model theoretic theorems describing the intersection of subvarieties of abelian varieties with certain special subgroups, we study intersections of varieties with certain uniformly definable groups containing the torsion groups in lieu of directly analyzing the torsion groups themselves. Unlike some of the other work, rather than offering an alternative proof of the Manin–Mumford conjecture itself, we use Raynaud's theorem to derive this uniform version.

Before recalling Buium's construction of p-differential characters on abelian varieties we highlight the crucial features of the groups obtained as the kernels of his characters that we shall exploit.

Definition 4.4. Let k be an algebraically closed field of characteristic $p > 0$. If X is a scheme over $W[k]$ and n is a natural number, then we write $\rho_n \colon X(W[k]) \to X(W[k]/p^{n+1}W[k])$ for the reduction modulo p^{n+1} map. If $n < m$, then we write $\rho_{m,n} \colon X(W[k]/p^{m+1}W[k]) \to X(W[k]/p^{n+1}W[k])$ for the intermediate reduction map. Using the Greenberg transform, $\rho_{m,n}$ may be regarded as a map of schemes over k. If $Z \subseteq X(W[k])$ is a closed subset of $X(W[k])$, then we say that Z is *finite dimensional* if for every natural number n the set $\rho_n(Z)$ is the set of k-rational points on a subvariety of $\rho_n(X(W[k])$ with respect to the identification $X(W[k]/p^{n+1}W[k])$ with the k-rational points of an algebraic variety over k and $\limsup \deg \rho_{n+1,n} \restriction (\rho_{n+1}Z)$ is finite.

At least when the characteristic of k is not two, Buium establishes that in analogy to the group homomorphisms constructed by Manin using derivations on function fields that if A is an abelian scheme over $W[k]$ of relative dimension g, then there is a group homomorphism given by a p-differential function $\mu \colon A(W[k]) \to W[k]^g$ for which the kernel, $A^\sharp(W[k])$, is finite dimensional. While the actual $A^\sharp$ groups need not vary uniformly, Buium does observe with Remark (1) on page 327 of [6] that the data required to produce group homomorphisms with finite dimensional kernels is bounded uniformly. Let us reformulate his observation as a theorem.

Theorem 4.5 (Buium). *Let k be an algebraically closed field of characteristic $p > 2$. Suppose that S is a variety (reduced, integral scheme of finite type) over $W[k]$ and that $A \to S$ is an abelian scheme over S of relative dimension g. Then there is a p-differential function $\mu \colon A(W[k]) \to W[k]^g$ such that for each $s \in S(W[k])$ the*

map $\mu_s\colon A_s(W[k]) \to W[k]^g$ is a group homomorphism for which $\ker(\mu_s)$ *is a finite dimensional proalgebraic group.*

Since the additive group is torsion free, the group $\ker(\mu_s)$ contains the torsion group $A_s(W[k])_{\mathrm{tor}}$. Moreover, since μ is a p-differential function, the group $\ker(\mu_s)$ is definable in $\mathcal{L}(\mathcal{A}, \sigma)$. In the notation of Theorem 4.1, one might like to argue that there are boundedly many points in $(X_s(W[k]) \smallsetminus \mathrm{Ueno}(X_s)(W[k])) \cap \ker(\mu_s)$ and then conclude *a fortiori* that the same is true with $\ker(\mu_s)$ replaced by the torsion subgroup of $A_s(W[k])$. Unfortunately, this stronger assertion is false in general. However, we shall establish a weak form of this boundedness statement for finite dimensional subgroups of abelian varieties from which Theorem 4.1 follows.

Theorem 4.6. *Let k be an algebraically closed field of characteristic p. Suppose that A is an abelian scheme over $W[k]$. Suppose $G \leq A(W[k])$ is a finite dimensional $\mathcal{L}(\mathcal{A}, \sigma)$-definable subgroup of $A(W[k])$. If $X \subseteq A$ is a closed subscheme, then $\rho_0((X(W[k]) \smallsetminus \mathrm{Ueno}(X)(W[k])) \cap G)$ is finite.*

Proof. If there were a counter-example to this theorem, then one could be found with $k = \mathbb{F}_p^{\mathrm{alg}}$. Indeed, by the quantifier elimination part of Theorem 3.4 the set $\rho_0((X(W[k]) \smallsetminus \mathrm{Ueno}(X)(W[k])) \cap G)$ is constructible. Hence, if it is infinite it contains a component of the form $Y(k) \smallsetminus F(k)$ where Y is an irreducible variety over k of dimension at least one and F is a proper subvariety. Since the extension $W[\mathbb{F}_p^{\mathrm{alg}}] \hookrightarrow W[k]$ is elementary, the assertion that there exist the appropriate parameters to define such an A, X, G, Y, and F is true in $W[\mathbb{F}_p^{\mathrm{alg}}]$. Likewise, if k' is an algebraically closed field of characteristic p, then because $W[\mathbb{F}_p^{\mathrm{alg}}] \hookrightarrow W[k']$ is elementary, we may transfer the counterexample from $W[\mathbb{F}_p^{\mathrm{alg}}]$ to $W[k']$. Thus, we may take k to be any algebraically closed field of characteristic p.

Let Y be as in the previous paragraph. Let $Z \subseteq Y$ be a curve with $Z(k) \cap F(k)$ finite. Translating, we may assume that Z contains the origin. Let H be the algebraic group generated by Z and let $\tilde{H} := (\rho_0^{-1} H(k)) \cap G$. Then $\tilde{H}$ is a definable, finite dimensional group for which $\rho_0((X(W[k]) \smallsetminus \mathrm{Ueno}(X)(W[k])) \cap \tilde{H})$ is infinite. Thus, we may and do assume that $G = \tilde{H}$.

We now transpose the proof of Proposition 4.4 of [12] to our unstable situation. For the moment we make use of our flexibility in the choice of k by taking k to be an algebraically closed field of characteristic p and cardinality strictly greater than that of the continuum. For each definable set $T \subseteq \ker(\rho_0 \restriction G)$, let $R_T := \{x \in Z(k) \mid (\exists g \in G)\ g + T = ((X(W[k]) \smallsetminus \mathrm{Ueno}(X)(W[k])) \cap G)_x\}$. The set R_T is a definable subset of the k-rational points of the curve Z and is thus either finite or cofinite. As $Z(k) = \bigcup_T R_T$ and there are at most continuum many such T and $|Z(k)| > 2^{\aleph_0}$, there must be some T for which R_T is cofinite. Translating T within $\ker(\rho_0 \restriction G)$, we may assume that T contains the origin. Let $S := \{x \in \ker(\rho_0 \restriction G) \mid x + T = T\}$. If $g + T$ is a fibre of $((X(W[k]) \smallsetminus \mathrm{Ueno}(X)(W[k])) \cap G)$, then $g + \overline{S} \subseteq X$ showing that g belongs to the Ueno locus of X unless S is finite, but g does not belong to the

Ueno locus of X. Thus, S must be finite. Thus, the correspondence which associates to $x \in Z(k)$ the g for which $g + T = ((X(W[k]) \setminus \mathrm{Ueno}(X)(W[k])) \cap G)_x$ is one to finite. Let $\tilde{Z}$ be the image of this correspondence in G. Note that $\tilde{Z}$ is a subset of $((X(W[k]) \setminus \mathrm{Ueno}(X)(W[k])) \cap G)$.

As the restriction of the map ρ_0 to $\tilde{Z}$ is finite to one, for $n \gg 0$ the map $\rho_{n+1,n}\colon \rho_{n+1}(\tilde{Z}) \to \rho_n(\tilde{Z})$ is a bijective morphism. Thus, we can find finitely many definable subsets $\tilde{Z}_1, \ldots, \tilde{Z}_m$ of $\tilde{Z}$ for which $\rho_n(\tilde{Z}_i)$ is always irreducible. For each such "component" if we translate $\tilde{Z}_i$ so that it contains the origin and then form the group L_i that it generates, we see that L_i is definable. Indeed, by the finite dimensionality of G the constructible sets $\rho_n(\tilde{Z}_i)$ generate an algebraic subgroup of $\rho_n(G)$ in a bounded number of steps. As the map ρ_0 is finite to one on $\tilde{Z}_i$, the same is true on L_i.

Now we use our flexibility in the choice of k to make k small: if $k = \mathbb{F}_p^{\mathrm{alg}}$, then every element of $\rho_0(L_i)$ is torsion. As the kernel of ρ_0 on L_i is finite, it follows that every element of L_i is torsion. By Raynaud's theorem, $L_i \cap ((X(W[k]) \setminus \mathrm{Ueno}(X)(W[k]))$ is finite. As this is true for each i, we conclude that the curve Z in question does not actually exist and that $\rho_0(((X(W[k]) \setminus \mathrm{Ueno}(X)(W[k])) \cap G))$ is finite after all. □

We are now in a position to complete the proof of Theorem 4.1.

Proof. Let $\mu\colon A(W[k]) \to W[k]^g$ be the p-differential function given by Theorem 4.5. By Theorem 4.6 each of the sets $\rho_0((X_s(W[k]) \setminus \mathrm{Ueno}(X_s)(W[k])) \cap \ker(\mu_s))$ is finite. By the quantifier elimination part of Theorem 3.4, this family of finite sets which *prima facie* is uniformly definable only in $W[k]$ is, in fact, uniformly definable in k. The quantifier "there exists infinitely many" may be eliminated in algebraically closed fields. Thus, there is a number B for which each of the above finite sets has cardinality at most B. Thus, the torsion points on X_s but outside the Ueno locus are contained in at most B cosets of the kernel of reduction. There is a bound $M = M(g, p)$ on the number of unramified torsion points in the kernel of reduction on an abelian scheme of relative dimension g depending just on g and p. Thus, there are at most $N := M \cdot B$ torsion points of $A_s(W[k])$ on X_s but outside the Ueno locus. □

References

[1] Ax, J., Kochen, S., Diophantine problems over local fields. I. *Amer. J. Math.* **87** (1965), 605–630.

[2] Ax, J., Kochen, S., Diophantine problems over local fields. II. A complete set of axioms for p-adic number theory. *Amer. J. Math.* **87** (1965), 631–648.

[3] Ax, J., Kochen, S., Diophantine problems over local fields. III. Decidable fields. *Ann. of Math.* (2) **83** (1966), 437–456.

[4] Basarab, Ş., Kuhlmann, F.-V., An isomorphism theorem for Henselian algebraic extensions of valued fields. *Manuscripta Math.* **77** (2–3) (1992), 113–126.

[5] Bélair, L., Macintyre, A., Scanlon, T., Model theory of Frobenius on Witt vectors. Preprint, 2002.

[6] Buium, A., Differential characters of abelian varieties over p-adic fields. *Invent. Math.* **122** (2) (1995), 309–340.

[7] Chatzidakis, Z., Hrushovski, E., Model theory of difference fields. *Trans. Amer. Math. Soc.* **351** (8) (1999), 2997–3071.

[8] Chatzidakis, Z., Hrushovski, E., Peterzil, Y., Model theory of difference fields. II. Periodic ideals and the trichotomy in all characteristics. *Proc. London Math. Soc.* (3) **85** (2) (2002), 257–311.

[9] Denef, J., van den Dries, L., p-adic and real subanalytic sets. *Ann. of Math.* (2) **128** (1) (1988), 79–138.

[10] Hazewinkel, M., *Formal groups and applications*. Pure Appl. Math. 78, Academic Press, Inc., New York, London 1978.

[11] Hodges, W., *Model Theory*. Encyclopedia Math. Appl. 42, Cambridge University Press, Cambridge 1993.

[12] Hrushovski, E., The Mordell-Lang conjecture for function fields. *J. Amer. Math. Soc.* **9** (3) (1996), 667–690.

[13] Hrushovski, E., Proof of Manin's theorem by reduction to positive characteristic. In *Model theory and algebraic geometry*, Lecture Notes in Math. 1696, Springer-Verlag, Berlin 1998, 197–205.

[14] Kuhlmann, F.-V., Quantifier elimination for Henselian fields relative to additive and multiplicative congruences. *Israel J. Math.* **85** (1–3) (1994), 277–306.

[15] Lipshitz, L., Robinson, Z., Uniform properties of rigid subanalytic sets. *Trans. Amer. Math. Soc.* **357** (11) (2005), 4349–4377.

[16] Raynaud, M., Sous-variétés d'une variété abélienne et points de torsion. In *Arithmetic and geometry*, Vol. I, Progr. Math. 35, Birkhäuser Boston, Boston, MA, 1983, 327–352.

[17] Scanlon, T., Quantifier elimination for the relative Frobenius. In *Valuation theory and its applications* (Saskatoon, SK, 1999), Vol. II, Fields Inst. Commun. 33, Amer. Math. Soc., Providence, RI, 2003, 323–352.

[18] Scanlon, T., Local André-Oort conjecture for the universal abelian variety. *Invent. Math.* **163** (1) (2006), 191–211.

[19] van den Dries, L., Analytic Ax-Kochen-Ersov theorems. In *Proceedings of the International Conference on Algebra* (Novosibirsk, 1989), Part 3, Contemp. Math. 131, Amer. Math. Soc., Providence, RI, 1992, 379–398.

[20] van den Dries, L., Haskell, D., Macpherson, D., One-dimensional p-adic subanalytic sets. *J. London Math. Soc.* (2) **59** (1) (1999), 1–20.

University of California, Berkeley, Department of Mathematics, Evans Hall, Berkeley, CA 94720-3840, U.S.A.

E-mail: scanlon@math.berkeley.edu

Borel superrigidity and the classification problem for the torsion-free abelian groups of finite rank

Simon Thomas*

Abstract. In 1937, Baer solved the classification problem for the torsion-free abelian groups of rank 1. Since then, despite the efforts of many mathematicians, no satisfactory solution has been found of the classification problem for the torsion-free abelian groups of rank $n \geq 2$. So it is natural to ask whether the classification problem for the higher rank groups is genuinely difficult. In this article, I will explain how this question can be partially answered, using ideas from descriptive set theory and Zimmer's superrigidity theory.

Mathematics Subject Classification (2000). Primary 03E15, 20K15, 37A20.

Keywords. Borel equivalence relation, superrigidity, torsion-free abelian group.

1. Introduction

In this article, we shall discuss some recent work which partially explains why no satisfactory system of complete invariants has yet been found for the torsion-free abelian groups of finite rank $n \geq 2$. Recall that, up to isomorphism, the torsion-free abelian groups A of rank n are exactly the additive subgroups of the n-dimensional vector space $\mathbb{Q}^n$ which contain n linearly independent elements. Thus the classification problem for the torsion-free abelian groups of rank n can be naturally identified with the corresponding problem for

$$R(\mathbb{Q}^n) = \{A \leqslant \mathbb{Q}^n \mid A \text{ contains } n \text{ linearly independent elements}\}.$$

In 1937, Baer [3] solved the classification problem for the class $R(\mathbb{Q})$ of torsion-free abelian groups of rank 1 as follows. Let $\mathbb{P}$ be the set of primes. Suppose that $G \in R(\mathbb{Q})$ and that $0 \neq x \in G$. Then for each $p \in \mathbb{P}$, the *p-height* of x is defined to be

$$h_x(p) = \sup\{n \in \mathbb{N} \mid \text{ There exists } y \in G \text{ such that } p^n y = x\} \in \mathbb{N} \cup \{\infty\};$$

and the *characteristic* $\chi(x)$ of x is defined to be the sequence

$$\langle h_x(p) \mid p \in \mathbb{P}\rangle \in (\mathbb{N} \cup \{\infty\})^{\mathbb{P}}.$$

*Research partially supported by NSF Grants.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Two sequences $\chi_1, \chi_2 \in (\mathbb{N} \cup \{\infty\})^{\mathbb{P}}$ are said to *belong to the same type*, written $\chi_1 \equiv \chi_2$, iff

(a) $\chi_1(p) = \chi_2(p)$ for almost all primes p; and

(b) if $\chi_1(p) \neq \chi_2(p)$, then both $\chi_1(p)$ and $\chi_2(p)$ are finite.

Clearly $\equiv$ is an equivalence relation on $(\mathbb{N} \cup \{\infty\})^{\mathbb{P}}$. Furthermore, it is easily checked that if $G \in R(\mathbb{Q})$, then $\chi(x) \equiv \chi(y)$ for all $0 \neq x, y \in G$. Hence we can define the *type* $\tau(G)$ of G to be the $\equiv$-equivalence class containing $\chi(x)$, where x is any non-zero element of G. In [3], Baer proved that $\tau(G)$ is a complete invariant for the isomorphism problem for the rank 1 groups.

Theorem 1.1 (Baer [3]). *If $G, H \in R(\mathbb{Q})$, then $G \cong H$ iff $\tau(G) = \tau(H)$.*

However, the situation is much less satisfactory in the case of the torsion-free abelian groups of rank $n \geq 2$. In the late 1930s, Kurosh [22] and Malcev [25] found complete invariants for these groups consisting of equivalence classes of infinite sequences $\langle M_p \mid p \in \mathbb{P} \rangle$ of matrices, where each $M_p \in \mathrm{GL}_n(\mathbb{Q}_p)$. Unfortunately, as Fuchs [8] remarks in his classic textbook, the associated equivalence relation is so complicated that the problem of deciding whether two sequences are equivalent is as difficult as that of deciding whether the corresponding groups are isomorphic. It is natural to ask whether the classification problem for the higher rank groups is genuinely more difficult than that for the rank 1 groups. Of course, if we wish to show that the classification problem for the groups of rank $n \geq 2$ is intractible, it is not enough merely to prove that there are $2^{\aleph_0}$ such groups up to isomorphism: for there are $2^{\aleph_0}$ pairwise nonisomorphic groups of rank 1 and we have just seen that Baer has given a satisfactory classification for this class of groups. In this article, following Friedman-Stanley [7] and Hjorth–Kechris [14], we shall explain how to use the more sensitive notions of descriptive set theory to measure the complexity of the classification problem for the groups of rank $n \geq 2$.

The basic idea is quite simple; namely, in order to understand the relative complexity of these and other classification problems, we shall consider the question of when one classification problem can be "explicitly reduced" to another. For example, the classification problem for the rank n groups can be explicitly reduced to that for the rank $n + 1$ groups by the map

$$\begin{aligned} R(\mathbb{Q}^n) &\to R(\mathbb{Q}^{n+1}) \\ A &\mapsto A \oplus \mathbb{Q} \end{aligned}$$

in the sense that

$$A \cong B \quad \text{iff} \quad A \oplus \mathbb{Q} \cong B \oplus \mathbb{Q}.$$

Of course, this observation is neither surprising nor particularly interesting; and we shall be more concerned with the question of whether there exists an "explicit map" in the opposite direction

$$f\colon R(\mathbb{Q}^{n+1}) \to R(\mathbb{Q}^n)$$

such that

$$A \cong B \quad \text{iff} \quad f(A) \cong f(B).$$

If we drop the requirement that f should be "explicit", then such a map certainly exists: since $R(\mathbb{Q}^{n+1})$ and $R(\mathbb{Q}^n)$ both contain $2^{\aleph_0}$ groups up to isomorphism, we can simply use the Axiom of Choice to match up the isomorphism classes. However, nobody would regard such a matching as a satisfactory reduction of one classification problem to another. In order to give a precise formulation of this question, it is first necessary to discuss some of the basic notions from the theory of Borel equivalence relations.

Let $(X, \mathcal{S})$ be a measurable space; i.e. a set X equipped with a σ-algebra $\mathcal{S}$ of subsets of X. Then $(X, \mathcal{S})$ is said to be a *standard Borel space* iff there exists a complete separable metric d on X such that $\mathcal{S}$ is the σ-algebra of Borel sets of (X, d). By a classic result of Kuratowski [21], if $(X, \mathcal{S})$ is an uncountable standard Borel space, then $(X, \mathcal{S})$ is measurably isomorphic to the unit interval $[0, 1]$ equipped with its σ-algebra of Borel sets. The obvious examples of standard Borel spaces include $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Q}_p$, as well as the Cantor space

$$2^C = \{h \mid h\colon C \to 2\},$$

where C is any countably infinite set. Furthermore, identifying each subset $B \subseteq C$ with its characteristic function $\chi_B \in 2^C$, it follows that the power set $\mathcal{P}(C)$ is also a standard Borel space. Less obviously, there is a uniform way to represent classes of countable structures, such as groups, fields, graphs, etc., by the elements of suitable standard Borel spaces. For example, in order to define the standard Borel space of countable graphs, we first restrict our attention to the set $\mathcal{C}$ of graphs

$$\Gamma = \langle \mathbb{N}, E_\Gamma \rangle$$

with vertex set $\mathbb{N}$. After identifying each such graph $\Gamma \in \mathcal{C}$ with its edge relation $E_\Gamma \in \mathcal{P}(\mathbb{N} \times \mathbb{N})$, it is easily checked that $\mathcal{C}$ is a Borel subset of the standard Borel space $\mathcal{P}(\mathbb{N} \times \mathbb{N})$; and this implies that $\mathcal{C}$ is also a standard Borel space. (For example, see Kechris [20].) It should be relatively clear how to generalise the method of this example to deal with other classes of countable structures. However, in this article, we shall mainly be concerned with the classes of torsion-free abelian groups of rank $n \geq 1$ and the class of finitely generated groups; and these classes can be more conveniently represented by the following more *ad hoc* spaces.

Example 1.2. Let $n \geq 1$. Then $R(\mathbb{Q}^n)$ is a Borel subset of the standard Borel space $\mathcal{P}(\mathbb{Q}^n)$ and so $R(\mathbb{Q}^n)$ is a standard Borel space. For later use, note that the natural action of $\mathrm{GL}_n(\mathbb{Q})$ on the vector space $\mathbb{Q}^n$ induces a corresponding action on $R(\mathbb{Q}^n)$; and that if $A, B \in R(\mathbb{Q}^n)$, then $A \cong B$ iff there exists an element $\varphi \in \mathrm{GL}_n(\mathbb{Q})$ such that $\varphi[A] = B$.

Example 1.3 (Champetier [4]). The standard Borel space $\mathcal{G}$ of finitely generated groups can be defined as follows. Let F be the free group on countably many generators $X = \{x_i \mid i \in \mathbb{N}\}$. Suppose that G is a finitely generated group and that $(g_0, \ldots, g_n)$ is a finite sequence of generators. Then, by considering the homomorphism $\pi : F \to G$ defined by

$$\pi(x_i) = \begin{cases} g_i & \text{if } 0 \leq i \leq n, \\ 1 & \text{otherwise,} \end{cases}$$

we see that G can be realized as a quotient F/N, where N is a normal subgroup which contains all but finitely many elements of the basis X. (Of course, choosing a different generating sequence usually results in a different realization.) Thus we can identify $\mathcal{G}$ with the set of all such normal subgroups N of F. With this identification, $\mathcal{G}$ is a Borel subset of the standard Borel space $\mathcal{P}(F)$ and hence $\mathcal{G}$ is a standard Borel space. As in Example 1.2, the isomorphism relation on the standard Borel space of finitely generated groups is the orbit equivalence relation of a natural action of a suitable countable group. More precisely, let $\mathrm{Aut}_f(F)$ be the subgroup of $\mathrm{Aut}(F)$ generated by the elementary Nielsen transformations

$$\{\alpha_i \mid i \in \mathbb{N}\} \cup \{\beta_{ij} \mid i \neq j \in \mathbb{N}\},$$

where α_i is the automorphism sending x_i to x_i^{-1} and leaving $X \setminus \{x_i\}$ fixed; and β_{ij} is the automorphism sending x_i to $x_i x_j$ and leaving $X \setminus \{x_i\}$ fixed. Then the natural action of $\mathrm{Aut}_f(F)$ on F induces a corresponding action on the space $\mathcal{G}$ of normal subgroups of F which contain all but finitely many elements of the basis X; and if $N, M \in \mathcal{G}$ are two such normal subgroups, then $F/N \cong F/M$ iff there exists $\varphi \in \mathrm{Aut}_f(F)$ such that $\varphi[N] = M$. (For example, see Champetier [4] and Lyndon–Schupp [24].)

If X, Y are standard Borel spaces, then $f : X \to Y$ is a *Borel map* iff $f^{-1}(B)$ is Borel for every Borel subset $B \subseteq Y$. Equivalently, f is Borel iff $\mathrm{graph}(f)$ is a Borel subset of $X \times Y$. Now suppose that E, F are equivalence relations on the standard Borel spaces X, Y respectively. (For example, X and Y could be spaces of countable structures and E, F could be the corresponding isomorphism relations.) Then E is *Borel reducible* to F, written $E \leq_B F$, if there exists a Borel map $f : X \to Y$ such that

$$x E y \quad \text{iff} \quad f(x) F f(y).$$

E and F are *Borel bireducible*, written $E \sim_B F$, if both $E \leq_B F$ and $F \leq_B E$. Finally we write $E <_B F$ if both $E \leq_B F$ and $F \not\leq_B E$.

Remark 1.4. Of course, the notion of a Borel reduction $f : X \to Y$ from E to F is intended to capture the intuitive idea of an "explicit reduction" from the E-classification problem to the F-classification problem. For example, with a little practice, it is easily checked that any given explicit map $f : \mathbb{R} \to \mathbb{R}$ is Borel. On the other hand, many mathematicians are reluctant to accept that an arbitrary Borel map $f : \mathbb{R} \to \mathbb{R}$ should

be regarded as explicit. However, it is not necessary for us to address this question, since we will mainly be concerned with non-reducibility results; and for such results, it is clearly preferable to work with the broadest possible class of maps. (It is perhaps worth mentioning that the proofs of our main results actually show that there are no measurable reductions between the relevant classification problems. By a well-known theorem of Solovay [30], the existence of a non-measurable map requires an essential use of the Axiom of Choice and so such maps are certainly not explicit.)

Example 1.5. For each $n \geq 1$, let $\cong_n$ be the isomorphism relation on $R(\mathbb{Q}^n)$. Then the map

$$\begin{aligned} R(\mathbb{Q}^n) &\to R(\mathbb{Q}^{n+1}), \\ A &\mapsto A \oplus \mathbb{Q}, \end{aligned}$$

is a Borel reduction from $\cong_n$ to $\cong_{n+1}$. Hence

$$(\cong_1) \leq_B (\cong_2) \leq_B \cdots \leq_B (\cong_n) \leq_B \cdots$$

and our earlier question of whether the classification problem for the higher rank groups is genuinely more difficult than that for the rank 1 groups can be interpreted as the question of whether $(\cong_1) <_B (\cong_2)$.

Before discussing the solution of this problem, it will be helpful to give a brief account of some of the theory of countable Borel equivalence relations. (A detailed development of this theory can be found in Jackson–Kechris–Louveau [17].) If X is a standard Borel space, then a *Borel equivalence relation* on X is an equivalence relation $E \subseteq X^2$ which is a Borel subset of X^2. The Borel equivalence relation E is said to be *countable* iff every E-equivalence class is countable. Most of the Borel equivalence relations that we shall consider in this article arise from group actions as follows. Let G be an *lcsc* group; i.e. a locally compact second countable group. Then a *standard Borel G-space* is a standard Borel space X equipped with a Borel action $(g, x) \mapsto g \cdot x$ of G on X. The corresponding G-orbit equivalence relation on X, which we shall denote by E_G^X, is a Borel equivalence relation. In fact, by Kechris [19], E_G^X is Borel bireducible with a countable Borel equivalence relation. Conversely, by Feldman–Moore [6], if E is an arbitrary countable Borel equivalence relation on the standard Borel space X, then there exists a countable group G and a Borel action of G on X such that $E = E_G^X$.

Example 1.6. As we pointed out in Examples 1.2 and 1.3, the isomorphism relations on the spaces $R(\mathbb{Q}^n)$ of torsion-free abelian groups of rank n and the space $\mathcal{G}$ of finitely generated groups are the orbit equivalence relations of natural actions of suitable countable groups. These actions are easily seen to be Borel and so each of these isomorphism relations is a countable Borel equivalence relation.

With respect to Borel reducibility, the least complex countable Borel equivalence relations are those which are *smooth*; i.e. those countable Borel equivalence relations E on a standard Borel space X for which there exists a Borel function $f : X \to Y$

into a standard Borel space Y such that $x\,E\,y$ iff $f(x) = f(y)$. Equivalently, the countable Borel equivalence relation E on X is smooth iff the quotient X/E is a standard Borel space. (Here X/E denotes the set of E-classes equipped with the quotient Borel structure.)

Example 1.7. The isomorphism relation on the standard Borel space of countable divisible abelian groups is smooth. To see this, recall that if A is a countable divisible abelian group, then $A = D \oplus T$, where T is the torsion subgroup and D is torsion-free. Let $r_0(A) \in \mathbb{N} \cup \{\infty\}$ be the rank of D; and for each prime p, let $r_p(A) \in \mathbb{N} \cup \{\infty\}$ be the rank of the p-component T_p of T. Then the invariant

$$\rho(A) = (r_0(A), r_2(A), r_3(A), \ldots, r_p(A), \ldots)$$

determines A up to isomorphism.

Next in complexity come those countable Borel equivalence relations E which are Borel bireducible with the *Vitali equivalence relation* E_0 defined on $2^{\mathbb{N}}$ by $x\,E_0\,y$ iff $x(n) = y(n)$ for almost all n. More precisely, by Harrington–Kechris–Louveau [12], if E is a countable Borel equivalence relation, then E is nonsmooth iff $E_0 \leq_B E$. Furthermore, by Dougherty–Jackson–Kechris [5], if E is a countable Borel equivalence relation on a standard Borel space X, then the following three properties are equivalent:

(1) $E \leq_B E_0$.

(2) E is *hyperfinite*; i.e. there exists an increasing sequence

$$F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n \subseteq \cdots$$

of finite Borel equivalence relations on X such that $E = \bigcup_{n \in \mathbb{N}} F_n$. (Here an equivalence relation F is said to be *finite* iff every F-equivalence class is finite.)

(3) There exists a Borel action of $\mathbb{Z}$ on X such that $E = E^X_{\mathbb{Z}}$.

Example 1.8. As is easily checked, Baer's classification of the rank 1 groups implies that $(\cong_1) \sim_B E_0$.

It turns out that there is also a most complex countable Borel equivalence relation E_∞, which is *universal* in the sense that $F \leq_B E_\infty$ for every countable Borel equivalence relation F; and, furthermore, $E_0 <_B E_\infty$. (Clearly this universality property uniquely determines E_∞ up to Borel bireducibility.) E_∞ has a number of natural realisations in many areas of mathematics, including algebra, topology and recursion theory. (See Jackson–Kechris–Louveau [17].) For example, E_∞ is Borel bireducible with both the isomorphism relation for finitely generated groups [36] and the isomorphism relation for fields of finite transcendence degree [37].

For many years, it was an open problem whether there existed infinitely many countable Borel equivalence relations E such that $E_0 <_B E <_B E_\infty$. This problem

was finally resolved by Adams–Kechris [2], who used Zimmer's superrigidity theory [38] to show that there are actually $2^{\aleph_0}$ such relations E up to Borel bireducibility. More recently, Hjorth–Kechris [15] have found an "elementary" proof of this result; i.e. a proof which requires no more background than the standard measure theory and functional analysis which should be known by every mathematician.

Returning to our discussion of the complexity of the isomorphism relation $\cong_n$ on the standard Borel space $R(\mathbb{Q}^n)$ of torsion-free abelian groups of rank n, we now see that

$$(\cong_1) \leq_B (\cong_2) \leq_B \cdots \leq_B (\cong_n) \leq_B \cdots \leq_B E_\infty.$$

In [14], Hjorth–Kechris conjectured that $(\cong_2) \sim_B E_\infty$; in other words, the classification problem for the torsion-free abelian groups of rank 2 is already as complex as that for arbitrary finitely generated groups. Of course, if true, this would have completely explained the failure to find a satisfactory system of complete invariants for the torsion-free abelian groups of rank $n \geq 2$, since nobody expects such a system to exist for the class of finitely generated groups. In [13], Hjorth provided some initial evidence for this conjecture by proving that the classification problem for the higher rank groups is indeed genuinely more difficult than that for the rank 1 groups.

Theorem 1.9 (Hjorth [13]). $(\cong_1) <_B (\cong_2)$.

However, the conjecture appeared considerably less plausible after Adams–Kechris [2] used Zimmer's superrigidity theory [38] to prove that

$$(\cong_1^*) <_B (\cong_2^*) <_B \cdots <_B (\cong_n^*) <_B \cdots$$

where $(\cong_n^*)$ is the restriction of the isomorphism relation to the class of *rigid* torsion-free abelian groups $A \in R(\mathbb{Q}^n)$. Here an abelian group A is said to be rigid if its only automorphisms are the obvious ones: $a \mapsto a$ and $a \mapsto -a$. In particular, it follows that none of the isomorphism relations $\cong_n^*$ is a universal countable Borel equivalence relation. Soon afterwards, making essential use of the earlier work of Hjorth [13] and Adams–Kechris [2], Thomas [31] proved the corresponding result for the isomorphism relation $\cong_n$ on the class $R(\mathbb{Q}^n)$ of all torsion-free abelian groups of rank n.

Theorem 1.10 (Thomas [31]). $(\cong_n) <_B (\cong_{n+1})$ *for all* $n \geq 2$.

Corollary 1.11. $(\cong_n) <_B (E_\infty)$ *for all* $n \geq 1$.

Unfortunately, while Theorem 1.10 shows that the relative complexity of the classification problem for the torsion-free abelian groups of rank n increases strictly with the rank n, it says little about the absolute complexity of these problems. In particular, it fails to answer the following:

Question 1.12. Is the classification problem for the torsion-free abelian groups of rank 2 "genuinely difficult"?

While it is difficult to imagine giving a precise formulation of Question 1.12, it certainly includes the question of whether $\cong_2$ is an immediate successor of $\cong_1$ with respect to Borel reducibility. In other words, does there exist a Borel equivalence relation E such that

$$(\cong_1) <_B E <_B (\cong_2)?$$

In seeking such an equivalence relation E, it is natural to consider the classification problem for various restricted classes of torsion-free abelian groups.

Definition 1.13. For each prime p and $n \geq 1$, let $R^p(\mathbb{Q}^n)$ be the standard Borel space of all p-local subgroups $A \leqslant \mathbb{Q}^n$ of rank n and let $\cong_n^p$ be the isomorphism relation on $R^p(\mathbb{Q}^n)$.

Here an abelian group A is said to be *p-local* iff A is q-divisible for all primes $q \neq p$. Of course, if an abelian group A is q-divisible for all primes q, then A is divisible and we have already seen that the divisible abelian groups are easily classified. Consequently, all of the complexity of the classification problem for the p-local groups is concentrated in the single prime p. In Thomas [31], it was shown that if the prime p is fixed, then

$$(\cong_1^p) <_B (\cong_2^p) <_B \cdots <_B (\cong_n^p) <_B \cdots .$$

But this left open the more natural question of whether the classification problem for the p-local torsion-free abelian groups of a *fixed* rank $n \geq 2$ was strictly easier than the classification problem for arbitrary torsion-free abelian groups of rank n. (It is trivial that $(\cong_1^p) <_B (\cong_1)$, since there are only two p-local groups of rank 1 up to isomorphism; namely, $\mathbb{Q}$ and $\mathbb{Z}_{(p)} = \{a/b \in \mathbb{Q} \mid b \text{ is relatively prime to } p\}$.) This question was partially answered in Thomas [33], where it was shown that if $n \geq 3$ and $p \neq q$ are distinct primes, then $\cong_n^p$ and $\cong_n^q$ are incomparable with respect to Borel reducibility. Of course, this implies that if $n \geq 3$, then $(\cong_n^p) <_B (\cong_n)$ for each prime p. Unfortunately, the argument in Thomas [33] made essential use of the fact that if $n \geq 3$, then $\mathrm{SL}_n(\mathbb{Z})$ is a Kazhdan group; and, consequently, the problem remained open when $n = 2$. This case was finally dealt with in Hjorth–Thomas [16], which ultimately depends upon the fact that $\mathrm{SL}_2(\mathbb{Z})$ satisfies a weak form of the Kazhdan property; namely, $\mathrm{SL}_2(\mathbb{Z})$ has Property (τ) with respect to its family of congruence subgroups. (For example, see Lubotzky [23].)

Theorem 1.14 (Hjorth–Thomas [16]). *If $p \neq q$ are distinct primes, then the classification problems for the p-local and q-local torsion-free abelian groups of rank 2 are incomparable with respect to Borel reducibility.*

Since it was already known [31] that $(\cong_1) <_B (\cong_2^p)$, it follows that

$$(\cong_1) <_B (\cong_2^p) <_B (\cong_2)$$

for each prime p; and hence there exists an infinite antichain of countable Borel equivalence relations which lie strictly between $(\cong_1)$ and $(\cong_2)$. With a little more effort,

it is possible to show that there are uncountably many countable Borel equivalence relations E such that

$$(\cong_1) <_B E <_B (\cong_2).$$

(It should be pointed out that the proof of the following result makes essential use of the work of Kurosh–Malcev [22], [25], which was so unfairly dismissed earlier in this section.)

Definition 1.15. If $P \subseteq \mathbb{P}$ is a set of primes, then an abelian group A is said to be *P-local* iff A is q-divisible for all primes $q \notin P$.

For example, an abelian group A is $\emptyset$-local iff A is divisible; while, on the other hand, every abelian group is $\mathbb{P}$-local. Clearly the class of P-local abelian groups is included in the class of Q-local groups iff $P \subseteq Q$.

Theorem 1.16 (Thomas [35]). *Let $n \geq 2$. If P, Q are sets of primes, then the classification problem for the P-local torsion-free abelian groups of rank n is Borel reducible to that for the Q-local groups of rank n iff $P \subseteq Q$.*

In particular, there exists an infinite chain $\{R_m \mid m \in \mathbb{N}\}$ of countable Borel equivalence relations such that

$$(\cong_1) <_B R_0 <_B R_1 <_B \cdots <_B R_m <_B \cdots <_B (\cong_2);$$

and so $\cong_2$ is very far from being an immediate successor of $\cong_1$ with respect to Borel reducibility.

Remark 1.17. It should be mentioned that $\mathrm{Id}_{\mathbb{R}}$, E_0 is the only known example (up to Borel bireducibility) of a pair of countable Borel equivalence relations E, F such that F is an immediate successor of E with respect to $\leq_B$. On the other hand, there are currently no countable Borel equivalence relations E with $E_0 \leq_B E <_B E_\infty$ for which it is known that no such countable Borel equivalence relation F exists.

2. Superrigidity

In this section, we shall discuss the orbit equivalence superrigidity theorems of Zimmer [38] and Furman [9], together with the corresponding Borel superrigidity theorems of Adams–Kechris [1], [2] and Thomas [32], [34]. Then, in the next section, we shall explain how to apply Borel superrigidity to the study of the classification problem for the torsion-free abelian groups of finite rank.

Recall that, by Feldman–Moore [6], if E is a countable Borel equivalence relation on the standard Borel space X, then there exists a countable group Γ and a Borel action of Γ on X such that $E = E_\Gamma^X$ is the corresponding orbit equivalence relation. However, it should be pointed out that the group Γ cannot be canonically recovered from E; and it is usually very difficult to determine whether two given Borel actions

of a pair Γ, Λ of countable groups give rise to Borel bireducible orbit equivalence relations. Consequently, the fundamental question in the study of countable Borel equivalence relations concerns the extent to which the data (X, E_Γ^X) determines the group Γ and its action on X. In order for there to be any chance of recovering Γ from this data, it is necessary to assume the following extra hypotheses:

(i) Γ acts *freely* on X; i.e. $\gamma \cdot x \neq x$ for all $1 \neq \gamma \in \Gamma$ and $x \in X$.

(ii) There exists a Γ-invariant probability measure μ on X.

For example, by Dougherty–Jackson–Kechris [5], if (i) holds and (ii) fails, then for any countable group $\Lambda \supseteq \Gamma$, there exists a free Borel action of Λ on X such that $E_\Lambda^X = E_\Gamma^X$. If Γ is finite, then E_Γ^X is smooth and so we shall suppose throughout this section that Γ is infinite. In this case, μ is necessarily nonatomic (i.e. $\mu(\{x\}) = 0$ for every $x \in X$) and it follows that the probability space (X, μ) is measurably isomorphic to the unit interval $[0, 1]$ equipped with its Lebesgue measure.

It is also natural to assume that the following "indecomposability hypothesis" holds:

(iii) Γ acts *ergodically* on (X, μ); i.e. every Γ-invariant Borel subset of X has measure 0 or 1.

Thus, even when working in the purely Borel setting, it is useful to focus our attention on those orbit equivalence relations which arise from free ergodic actions of countable groups on probability spaces.

Example 2.1. Let Γ be any countable group. Then the shift action of Γ on 2^Γ is defined by

$$(\gamma \cdot h)(\delta) = h(\gamma^{-1}\delta), \quad \gamma, \delta \in \Gamma,\ h \in 2^\Gamma.$$

Let μ be the usual product probability measure on 2^Γ. Then μ is Γ-invariant and Γ acts ergodically on $(2^\Gamma, \mu)$. (For example, see Hjorth–Kechris [15].) Furthermore, letting

$$(2)^\Gamma = \{h \in 2^\Gamma \mid \gamma \cdot h \neq h \text{ for all } 1 \neq \gamma \in \Gamma\}$$

be the free part of the action, it is easily checked that $\mu((2)^\Gamma) = 1$.

Now suppose that Γ, Λ are countable groups with free ergodic Borel actions on the probability spaces (X, μ), (Y, ν) respectively. Then, by Dougherty–Jackson–Kechris [5], the corresponding countable Borel equivalence relations E_Γ^X and E_Λ^Y are Borel bireducible iff there exist Borel complete sections $A \subseteq X$, $B \subseteq Y$ such that the restricted equivalence relations are isomorphic via a Borel bijection

$$f : (A,\ E_\Gamma^X \restriction A) \cong (B,\ E_\Lambda^Y \restriction B).$$

Here, for example, $A \subseteq X$ is said to be a *complete section* of E_Γ^X iff A intersects every E_Γ^X-class. In particular, it follows that $\mu(A), \nu(B) > 0$. However, there is no

reason to suppose that f preserves the corresponding "rescaled" probability measures μ_A, ν_B on A, B respectively, defined by $\mu_A(Z) = \mu(Z)/\mu(A)$, etc. If we add the requirement that the map f should also be measure-preserving, then we pass from the purely Borel setting into the richer measure-theoretic setting, where the fundamental question now concerns the extent to which the data (X, E_Γ^X, μ) determines the group Γ and its action on X.

Definition 2.2. With the above hypotheses, the actions of Γ, Λ on (X, μ), (Y, ν) are said to be *weakly orbit equivalent* iff there exist Borel subsets $A \subseteq X$, $B \subseteq Y$ with $\mu(A), \nu(B) > 0$ such that the restricted equivalence relations are isomorphic via a measure-preserving Borel bijection

$$f\colon (A,\ E_\Gamma^X \restriction A,\ \mu_A) \cong (B,\ E_\Lambda^Y \restriction B,\ \nu_B).$$

If $\mu(A) = \nu(B) = 1$, then the actions are said to be *orbit equivalent*.

Warning 2.3. At first glance, it might appear that weak orbit equivalence implies Borel bireducibility. However, this is not the case. In the measure-theoretic setting, sets and maps are only considered modulo measure zero sets; and, in particular, the Borel sets A, B in Definition 2.2 are not required to be complete sections.

Definition 2.4. The actions of Γ, Λ on (X, μ), (Y, ν) are said to be *isomorphic* iff there exist

- invariant Borel subsets $X_0 \subseteq X$, $Y_0 \subseteq Y$ with $\mu(X_0) = \nu(Y_0) = 1$,
- a measure-preserving Borel bijection $f\colon X_0 \to Y_0$, and
- a group isomorphism $\varphi\colon \Gamma \to \Lambda$

such that $f(\gamma \cdot x) = \varphi(\gamma) \cdot f(x)$ for all $\gamma \in \Gamma$ and $x \in X_0$.

If the actions of Γ, Λ on (X, μ), (Y, ν) are isomorphic, then they are clearly orbit equivalent. The strongest conceivable superrigidity theorem would say that, conversely, if the actions are (weakly) orbit equivalent, then they are necessarily isomorphic. Of course, in order for anything like this to be true, it is necessary to impose strong hypotheses on the groups involved. For example, Ornstein–Weiss [27] have shown that if Γ and Λ are amenable groups, then any free ergodic actions of Γ, Λ are orbit equivalent.

Definition 2.5. Let G be an lcsc group and let m be a fixed Haar measure on G. Then a subgroup $\Gamma \leqslant G$ is a *lattice* iff Γ is discrete and the covolume $m(G/\Gamma)$ is finite.

Suppose now that Γ is a lattice in a connected simple Lie group G such that $\mathbb{R}\text{-rank}(G) \geq 2$. For example, we can take $\Gamma = \mathrm{SL}_n(\mathbb{Z})$ and $G = \mathrm{SL}_n(\mathbb{R})$ for any $n \geq 3$. Then, while the lattice Γ is not uniquely determined by (X, E_Γ^X, μ), Zimmer's orbit equivalence superrigidity theorem says that this data does uniquely determine

the ambient Lie group G. More precisely, suppose that G_0 and G_1 are connected centerless simple Lie groups of $\mathbb{R}$-rank at least 2 and that Γ_0, Γ_1 are lattices in G_0, G_1 respectively. (In order to keep our account as transparent as possible, we shall mainly focus on the case of lattices in connected *centerless* simple Lie groups.) Suppose that Γ_0, Γ_1 have free ergodic Borel actions on the probability spaces (X_0, μ_0) and (X_1, μ_1). Then, for each $0 \leq i \leq 1$, there is a naturally associated induced action of G_i on the standard Borel space

$$\widehat{X}_i = X_i \times (G_i/\Gamma_i)$$

with invariant ergodic probability measure $\widehat{\mu}_i = \mu_i \times m_i$, where m_i is the Haar probability measure on G_i/Γ_i.

Theorem 2.6 (Zimmer [38]). *With the above hypotheses, if the actions of* Γ_0, Γ_1 *on* (X_0, μ_0), (X_1, μ_1) *are weakly orbit equivalent, then the induced actions of* G_0, G_1 *on* $(\widehat{X}_0, \widehat{\mu}_0)$, $(\widehat{X}_1, \widehat{\mu}_1)$ *are isomorphic. In particular,* $G_0 \cong G_1$.

Unfortunately, there are many examples of lattices Γ_0, Γ_1 with free ergodic Borel actions on probability spaces (X_0, μ_0), (X_1, μ_1) for which there exists a Borel reduction $f : X_0 \to X_1$ from $E_{\Gamma_0}^{X_0}$ to $E_{\Gamma_1}^{X_1}$ such that $\mu_1(f[X_0]) = 0$; and hence Zimmer's orbit equivalence superrigidity theorem cannot be directly applied in the purely Borel setting. However, it was an important insight of Adams–Kechris [2] that it is possible to apply Zimmer's more fundamental cocycle superrigidity theorem. (The notion of a cocycle will not be defined in this article. Clear accounts of the theory of cocycles can be found in Zimmer [38] and Adams–Kechris [2]. In particular, Adams–Kechris [2] provides a convenient introduction to the basic techniques and results in this area, written for the non-expert in the ergodic theory of groups.)

Theorem 2.7 (Adams–Kechris [2]). *With the above hypotheses, if* $E_{\Gamma_0}^{X_0} \leq_B E_{\Gamma_1}^{X_1}$, *then* G_0 *is involved in* G_1*; i.e. there exist Lie subgroups* $N \trianglelefteq H \leqslant G_1$ *such that* $G_0 \cong H/N$. *Consequently, if* $E_{\Gamma_0}^{X_0} \sim_B E_{\Gamma_1}^{X_1}$, *then* $G_0 \cong G_1$.

Corollary 2.8 (Adams–Kechris [2]). *There exist infinitely many countable Borel equivalence relations up to Borel bireducibility.*

In fact, by considering Borel actions of suitable S-arithmetic groups for various (possibly infinite) sets of primes S, Adams–Kechris [2] were able to prove that there are $2^{\aleph_0}$ such relations up to Borel bireducibility.

Corollary 2.9 (Adams–Kechris [2]). *There exist countable Borel equivalence relations which are incomparable with respect to Borel reducibility.*

The methods introduced by Adams–Kechris [2] are suitable for distinguishing between orbit equivalence relations of the form E_Γ^X and E_Λ^Y, where Γ and Λ are lattices in nonisogeneous higher rank semisimple Lie groups. More generally, they can be used to show that the countable Borel equivalence relations arising from suitably

chosen actions of "large" linear groups cannot be Borel reducible to those arising from the actions of "smaller" linear groups. For example, as we shall see in the next section, a variant of Theorem 2.7 can be used to prove that if $n \geq 2$, then the orbit equivalence relation arising from the action of $\mathrm{GL}_{n+1}(\mathbb{Q})$ on the standard Borel space $R(\mathbb{Q}^{n+1})$ of torsion-free abelian groups of rank $n+1$ is not Borel reducible to that arising from the action of $\mathrm{GL}_n(\mathbb{Q})$ on $R(\mathbb{Q}^n)$; in other words, $(\cong_{n+1}) \not\leq_B (\cong_n)$. Since we have already observed that $(\cong_n) \leq_B (\cong_{n+1})$, this implies that $(\cong_n) <_B (\cong_{n+1})$; i.e. that the complexity of the classification problem for the torsion-free abelian groups of rank n increases strictly with the rank n.

However, the methods of Adams–Kechris [2] are not as well-suited for those problems which involve distinguishing between orbit equivalence relations arising from different actions of the same countable group; e.g. the isomorphism relations for the p-local torsion-free abelian groups of rank 2, which arise as the orbit equivalence relations of the actions of $\mathrm{GL}_2(\mathbb{Q})$ on the standard Borel spaces $R^p(\mathbb{Q}^2)$. The next breakthrough occurred when Adams [1], by combining the use of Zimmer's cocycle superrigidity theorem with Ratner's measure classification theorem [29], developed a method for distinguishing between the orbit equivalence relations arising from suitably chosen actions of (not necessarily distinct) lattices Γ, Δ in the same higher rank semisimple Lie group G. (This idea had already been successfully exploited in the measure-theoretic setting by Zimmer [39] and Furman [9].) It quickly became clear that Adams' techniques were applicable to a wide range of natural problems concerning countable Borel equivalence relations. For example, combining the ideas of Adams [1] and Gefter–Golodets [10], it is straightforward to show that if $n \geq 3$, then the orbit equivalence relations arising from the following uncountable family of $\mathrm{SL}_n(\mathbb{Z})$-actions are pairwise incomparable with respect to Borel reducibility.

Example 2.10 (Gefter–Golodets [10]). Fix some integer $n \geq 2$ and for each nonempty set $\emptyset \neq J \subseteq \mathbb{P}$ of primes, let

$$K_n(J) = \prod_{p \in J} \mathrm{SL}_n(\mathbb{Z}_p),$$

where $\mathbb{Z}_p$ is the ring of p-adic integers. Then $K_n(J)$ is a compact group and we can regard $\mathrm{SL}_n(\mathbb{Z})$ as a subgroup of $K_n(J)$ via the diagonal embedding. Let μ_J be the Haar probability measure on $K_n(J)$ and let E_J be the orbit equivalence relation arising from the free action of $\mathrm{SL}_n(\mathbb{Z})$ on $K_n(J)$ via left translations. By the Strong Approximation Theorem [28], $\mathrm{SL}_n(\mathbb{Z})$ is a dense subgroup of $K_n(J)$ and this implies that $\mathrm{SL}_n(\mathbb{Z})$ acts ergodically on $(K_n(J), \mu_J)$.

Theorem 2.11 (Thomas [32]). *Fix some integer $n \geq 3$. If $J_0 \neq J_1$ are distinct nonempty subsets of $\mathbb{P}$, then E_{J_0} and E_{J_1} are incomparable with respect to Borel reducibility.*

The measure-theoretic analogue of this result was proved earlier by Gefter–Golodets [10], who showed that for distinct $J_0 \neq J_1$, the actions of $\mathrm{SL}_n(\mathbb{Z})$ on $(K_n(J_0), \mu_{J_0})$

and $(K_n(J_1), \mu_{J_1})$ are not weakly orbit equivalent. More recently, Furman [9] has shown that for many free ergodic actions of lattices Γ on probability spaces (X, μ), both the group Γ and its action on X are "almost uniquely determined" by the orbit equivalence relation E_Γ^X and the measure μ. More precisely, in our particular case, Furman's result takes the following form. (It is easily seen that if $J_0 \neq J_1$, then the actions of $\mathrm{SL}_n(\mathbb{Z})$ on $(K_n(J_0), \mu_{J_0})$ and $(K_n(J_1), \mu_{J_1})$ are not virtually isomorphic. Thus the following result is strictly stronger than that of Gefter–Golodets [10].)

Theorem 2.12 (Furman [9]). *Let $n \geq 3$ and let J be a nonempty subset of $\mathbb{P}$. Suppose that Λ is an* arbitrary *countable group with a free ergodic action on the probability space (Y, ν). If the actions of $\mathrm{SL}_n(\mathbb{Z})$, Λ on the probability spaces $(K_n(J), \mu_J)$, (Y, ν) are weakly orbit equivalent, then:*

(a) *$\mathrm{SL}_n(\mathbb{Z})$ and Λ are virtually isomorphic; and*

(b) *the actions of $\mathrm{SL}_n(\mathbb{Z})$, Λ on the probability spaces $(K_n(J), \mu_J)$, (Y, ν) are virtually isomorphic.*

Here two countable groups G_0, G_1 are said to be *virtually isomorphic* iff there exist subgroups $H_i \leqslant G_i$ of finite index and finite normal subgroups $N_i \trianglelefteq H_i$ for $i = 0, 1$ such that $H_0/N_0 \cong H_1/N_1$; and the free ergodic actions of G_0, G_1 on the probability spaces (X_0, μ_0), (X_1, μ_1) are said to be *virtually isomorphic* iff, after passing to ergodic components, the induced actions of H_0/N_0, H_1/N_1 on the factor spaces $(X_0, \mu_0)/N_0$, $(X_1, \mu_1)/N_1$ are isomorphic.

No analogues of Furman's results have yet been proved in the purely Borel setting, where all of the currently known superrigidity results impose very restrictive conditions on both the domain and the range of the relevant Borel bireduction. However, it seems reasonable to conjecture that the corresponding strengthening of Theorem 2.11 also holds in this setting.

Conjecture 2.13. The conclusion of Theorem 2.12 remains true if *weak orbit equivalence* is replaced by *Borel bireducibility*.

It is not known whether the analogue of Theorem 2.11 also holds when $n = 2$. Here the main obstacle is the failure of Zimmer's cocycle superrigidity theorem for the low rank Lie group $\mathrm{SL}_2(\mathbb{R})$. For the same reason, it is also not known whether or not these $\mathrm{SL}_2(\mathbb{Z})$-actions are (weakly) orbit equivalent. Since $\mathrm{SL}_2(\mathbb{Z})$ contains the free group F_2 on two generators as a subgroup of finite index, a positive solution of the following problem would also provide uncountably many "natural" F_2-actions which are pairwise neither Borel bireducible nor weakly orbit equivalent. Currently only three nonsmooth F_2-actions are known up to Borel bireducibility. On the other hand, in the measure-theoretic setting, Gaboriau–Popa [11] have recently constructed uncountably many F_2-actions which are pairwise not weakly orbit equivalent.

Conjecture 2.14. If $J_0 \neq J_1$ are distinct nonempty sets of primes, then the actions of $\mathrm{SL}_2(\mathbb{Z})$ on $(K_2(J_0), \mu_{J_0})$ and $(K_2(J_1), \mu_{J_1})$ are neither comparable with respect to Borel bireducibility nor weakly orbit equivalent.

We obtain a more manageable problem if we replace the lattice $\mathrm{SL}_2(\mathbb{Z})$ by

$$\Gamma_S = \mathrm{SL}_2(\mathbb{Z}[1/p_1, \ldots, 1/p_t]),$$

where $S = \{p_1, \ldots, p_t\}$ is a nonempty finite set of primes. Of course, Γ_S is no longer a lattice in $\mathrm{SL}_2(\mathbb{R})$. However, if we identify Γ_S with its image under the diagonal embedding into

$$G = \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{Q}_{p_1}) \times \cdots \times \mathrm{SL}_2(\mathbb{Q}_{p_t}),$$

then Γ_S is a lattice in G and Zimmer's cocycle superrigidity theorem holds for G. Furthermore, by Margulis–Tomanov [26], the analogue of Ratner's measure classification theorem also holds for G. For each nonempty (possibly infinite) set of primes J such that $S \cap J = \emptyset$, let E_S^J be the orbit equivalence relation arising from the action of Γ_S on

$$K_2(J) = \prod_{p \in J} \mathrm{SL}_2(\mathbb{Z}_p)$$

by left translations, where Γ_S is regarded as a subgroup of $K_2(J)$ via the diagonal embedding.

Theorem 2.15 (Thomas [34]). *Suppose that S_0, S_1 are nonempty finite sets of primes and that J_0, J_1 are nonempty (possibly infinite) sets of primes such that $S_0 \cap J_0 = S_1 \cap J_1 = \emptyset$. If $(J_0, S_0) \neq (J_1, S_1)$, then $E_{S_0}^{J_0}$ and $E_{S_1}^{J_1}$ are incomparable with respect to Borel reducibility.*

The proof of Theorem 2.15 easily extends to the more general situation of Γ_S-actions on homogeneous $K_2(J)$-spaces. For example, it is well-known that the compact group $\mathrm{SL}_2(\mathbb{Z}_p)$ acts transitively on the projective line $\mathbb{Q}_p \cup \{\infty\}$ over the field of p-adic numbers.

Theorem 2.16 (Thomas [34]). *Suppose that p, q are primes and that S, T are finite nonempty sets of primes such that $p \notin S$, $q \notin T$. If $(p, S) \neq (q, T)$, then the orbit equivalence relations of Γ_S, Γ_T on the projective lines $\mathbb{Q}_p \cup \{\infty\}$, $\mathbb{Q}_q \cup \{\infty\}$ are incomparable with respect to Borel reducibility.*

As we shall see in the next section, a variant of Theorem 2.16 can be used to prove that if $p \neq q$ are distinct primes, then the classification problems for the p-local and q-local torsion-free abelian groups of rank 2 are incomparable with respect to Borel reducibility.

3. The classification problem for the torsion-free abelian groups of finite rank

In this final section, we shall explain how to apply Borel superrigidity to the study of the classification problem for the torsion-free abelian groups of finite rank. First we shall

sketch the proof of Theorem 1.10, which says that the complexity of the isomorphism relation $\cong_n$ for the torsion-free abelian groups of rank n increases strictly with the rank n. This will be followed by a sketch of the proof of Theorem 1.14, which says that if $p \neq q$ are distinct primes, then the classification problems for the p-local and q-local torsion-free abelian groups of rank 2 are incomparable with respect to Borel reducibility.

Recall that for each $m \geq 1$, the isomorphism relation $\cong_m$ is precisely the orbit equivalence relation arising from the natural action of $\mathrm{GL}_m(\mathbb{Q})$ on the standard Borel space $R(\mathbb{Q}^m)$ of torsion-free abelian groups of rank m. In the last section, we saw that if Γ is a lattice in a higher rank centerless simple Lie group G and Γ has a free ergodic action on the probability space (X, μ), then the orbit equivalence relation E_Γ^X "encodes" the ambient Lie group G. More precisely, suppose that Λ is also a lattice in a centerless simple Lie group H and that Λ has a free ergodic action on the probability space (Y, ν). By Theorem 2.7, if $E_\Gamma^X \leq_B E_\Lambda^Y$, then G is involved in H; and, in particular, it follows that $\dim G \leq \dim H$. This certainly suggests that the orbit equivalence relation of $\mathrm{GL}_{n+1}(\mathbb{Q})$ on $R(\mathbb{Q}^{n+1})$ should not be Borel reducible to the orbit equivalence relation of $\mathrm{GL}_n(\mathbb{Q})$ on $R(\mathbb{Q}^n)$. Unfortunately, we cannot apply Theorem 2.7 directly to our situation, since:

(i) $\mathrm{GL}_m(\mathbb{Q})$ is not a lattice.

(ii) There does not exist a $\mathrm{GL}_m(\mathbb{Q})$-invariant probability measure on $R(\mathbb{Q}^m)$.

(iii) $\mathrm{GL}_m(\mathbb{Q})$ does not act freely on $R(\mathbb{Q}^m)$.

Fortunately, none of these difficulties is insurmountable. Suppose that $n \geq 2$ and that $f : R(\mathbb{Q}^{n+1}) \to R(\mathbb{Q}^n)$ is a Borel reduction from $\cong_{n+1}$ to $\cong_n$. First, following the example of Hjorth [13] and Adams–Kechris [2], we shall use the following result to deal with points (i) and (ii).

Theorem 3.1 (Hjorth [13]). *For each $m \geq 2$, there exists a nonatomic $\mathrm{SL}_m(\mathbb{Z})$-invariant ergodic probability measure μ on $R(\mathbb{Q}^m)$.*

In fact, Hjorth [13] has shown that for each prime $p \in \mathbb{P}$, there exists a nonatomic $\mathrm{SL}_m(\mathbb{Z})$-invariant ergodic probability measure μ_p on $R(\mathbb{Q}^m)$ which concentrates on the Borel subspace $R^p(\mathbb{Q}^m)$ of p-local groups. Later in this section, we shall sketch a proof of this result in the special case when $m = 2$.

Continuing the proof of Theorem 1.10, let E be the orbit equivalence relation arising from the action of the subgroup $\mathrm{SL}_{n+1}(\mathbb{Z})$ of $\mathrm{GL}_{n+1}(\mathbb{Q})$ on $R(\mathbb{Q}^{n+1})$. Then we can regard f as a countable-to-one Borel homomorphism from E to $\cong_n$; and Theorem 1.10 is an easy consequence of the following result. (As we shall see, most of our effort during the proof of Theorem 3.3 will go into dealing with point (iii).)

Definition 3.2. If E, F are equivalence relations on the standard Borel spaces X, Y, then the Borel map $f : X \to Y$ is a *Borel homomorphism* from E to F iff

$$x E y \text{ implies } f(x) F f(y) \quad \text{for all } x, y \in X.$$

Theorem 3.3 (Thomas [32]). *Let $m \geq 3$ and let X be a standard Borel $\mathrm{SL}_m(\mathbb{Z})$-space with an invariant ergodic probability measure μ. Suppose that $1 \leq n < m$ and that $f\colon X \to R(\mathbb{Q}^n)$ is a Borel homomorphism from $E^X_{\mathrm{SL}_m(\mathbb{Z})}$ to $\cong_n$. Then there exists an $\mathrm{SL}_m(\mathbb{Z})$-invariant Borel subset M with $\mu(M) = 1$ such that f maps M into a single $\cong_n$-class.*

Hence, letting μ be a nonatomic $\mathrm{SL}_{n+1}(\mathbb{Z})$-invariant ergodic probability measure on $R(\mathbb{Q}^{n+1})$, there exists an $\mathrm{SL}_{n+1}(\mathbb{Z})$-invariant Borel subset $M \subseteq R(\mathbb{Q}^{n+1})$ with $\mu(M) = 1$ such that f maps M into a single $\cong_n$-class $\mathcal{C}$. However, this is impossible, since $f^{-1}(\mathcal{C})$ consists of only countably many $\mathrm{SL}_{n+1}(\mathbb{Z})$-orbits. Hence $(\cong_n) <_B (\cong_{n+1})$ for all $n \geq 2$.

Next we shall sketch the proof of Theorem 3.3. Suppose that $m \geq 3$ and that X is a standard Borel $\mathrm{SL}_m(\mathbb{Z})$-space with an invariant ergodic probability measure μ. Suppose further that $1 \leq n < m$ and that $f\colon X \to R(\mathbb{Q}^n)$ is a Borel homomorphism from $E^X_{\mathrm{SL}_m(\mathbb{Z})}$ to $\cong_n$. We shall make use of the following variant of Theorem 2.7, which is a straightforward consequence of Zimmer's cocycle superrigidity theorem [38] and the ideas of Adams–Kechris [2].

Theorem 3.4 (Thomas [31]). *Let $m \geq 3$ and let X be a standard Borel $\mathrm{SL}_m(\mathbb{Z})$-space with an invariant ergodic probability measure μ. Suppose that $H \leqslant G(\mathbb{Q})$, where G is an algebraic $\mathbb{Q}$-group such that $\dim G < m^2 - 1$, and that H acts freely on the standard Borel H-space Y. If $f\colon X \to Y$ is a Borel homomorphism from $E^X_{\mathrm{SL}_m(\mathbb{Z})}$ to E^Y_H, then there exists an $\mathrm{SL}_m(\mathbb{Z})$-invariant Borel subset $M \subseteq X$ with $\mu(M) = 1$ such that f maps M into a single H-orbit.*

As we mentioned earlier, the action of $\mathrm{GL}_n(\mathbb{Q})$ on $R(\mathbb{Q}^n)$ is not free: in fact, for each $A \in R(\mathbb{Q}^n)$, the stabilizer of A in $\mathrm{GL}_n(\mathbb{Q})$ is precisely the automorphism group $\mathrm{Aut}(A)$ of A. Thus we are not yet in a position to apply Theorem 3.4.

Remark 3.5. This is actually a serious problem. The proof of Theorem 3.4 makes essential use of Zimmer's cocycle superrigidity theorem; and if H does not act freely on Y, then it impossible to define the associated cocycle on which the proof depends.

From now on, let $A_x = f(x) \in R(\mathbb{Q}^n)$. Roughly speaking, our strategy will be as follows. Suppose that there exists a Borel subset $X_0 \subseteq X$ with $\mu(X_0) = 1$ and a *fixed* subgroup $L \leqslant \mathrm{GL}_n(\mathbb{Q})$ such that $\mathrm{Aut}(A_x) = L$ for all $x \in X_0$. Then the equivalence relation $\cong_n \restriction f(X_0)$ will be induced by a free action of the quotient group $H = N_{\mathrm{GL}_n(\mathbb{Q})}(L)/L$ on the Borel subset

$$Y = \{A \in R(\mathbb{Q}^n) \mid \mathrm{Aut}(A) = L\}$$

of $R(\mathbb{Q}^n)$. Hence, provided that the quotient group H is isomorphic to a subgroup of an algebraic $\mathbb{Q}$-group $G(\mathbb{Q})$ with $\dim G < m^2 - 1$, we can apply Theorem 3.4. But why should X_0 and L exist? Imagine for the moment that there are only countably many possibilities for the subgroup $\mathrm{Aut}(A_x) \leqslant \mathrm{GL}_n(\mathbb{Q})$. Then there exists a Borel subset

$Z \subseteq X$ with $\mu(Z) > 0$ and a fixed subgroup $L \leqslant \mathrm{GL}_n(\mathbb{Q})$ such that $\mathrm{Aut}(A_x) = L$ for all $x \in Z$. Since $\mathrm{SL}_m(\mathbb{Z})$ acts ergodically on (X, μ), it follows that

$$X_0 = \{\gamma \cdot x \mid \gamma \in \mathrm{SL}_m(\mathbb{Z}) \text{ and } x \in Z\}.$$

has μ-measure 1. Let $g\colon X \to \mathrm{SL}_m(\mathbb{Z})$ be a Borel function such that $g(x) \cdot x \in Z$ for all $x \in X_0$. Then, replacing f by the Borel homomorphism f' defined by $f'(x) = f(g(x) \cdot x)$, we can suppose that $\mathrm{Aut}(A_x) = L$ for all $x \in X_0$.

Unfortunately, this approach does not work, since there are uncountably many possibilities for the subgroup $\mathrm{Aut}(A_x) \leqslant \mathrm{GL}_n(\mathbb{Q})$. In order to get around this difficulty, we shall shift our attention from the isomorphism relation on $R(\mathbb{Q}^n)$ to the coarser relation of quasi-isomorphism. This relation was first introduced in Jónsson [18], where it was shown that the class of torsion-free abelian groups of finite rank has a better decomposition theory with respect to quasi-isomorphism than with respect to isomorphism. This decomposition theory will not concern us in this article. Rather we shall exploit the fact that much of the number-theoretical complexity of finite rank torsion-free abelian groups is lost when they are only considered up to quasi-isomorphism; and this turns out to be enough to ensure that there are only countably many possibilities for the group of quasi-automorphisms of $A \in R(\mathbb{Q}^n)$.

Definition 3.6. If $A, B \in R(\mathbb{Q}^n)$, then A and B are said to be *quasi-equal*, written $A \approx_n B$, iff $A \cap B$ has finite index in both A and B.

Definition 3.7. If $A, B \in R(\mathbb{Q}^n)$, then A and B are said to be *quasi-isomorphic* iff there exists $\varphi \in \mathrm{GL}_n(\mathbb{Q})$ such that $\varphi[A] \approx_n B$.

It is easily checked that $\approx_n$ is a countable Borel equivalence relation on $R(\mathbb{Q}^n)$. For each $A \in R(\mathbb{Q}^n)$, let $[A]$ be the $\approx_n$-class containing A. We shall consider the induced action of $\mathrm{GL}_n(\mathbb{Q})$ on the set of $\approx_n$-classes. In order to describe the setwise stabilizer in $\mathrm{GL}_n(\mathbb{Q})$ of a $\approx_n$-class $[A]$, it is first necessary to introduce the notions of a quasi-endomorphism and a quasi-automorphism. If $A \in R(\mathbb{Q}^n)$, then a linear transformation $\varphi \in \mathrm{Mat}_n(\mathbb{Q})$ is said to be a *quasi-endomorphism* of A iff there exists an integer $m > 0$ such that $m\varphi[A] \leqslant A$. In other words, φ is a quasi-endomorphism of A iff there exists an integer $m > 0$ such that $m\varphi \in \mathrm{End}(A)$. It is easily checked that the collection $\mathrm{QE}(A)$ of quasi-endomorphisms of A is a $\mathbb{Q}$-subalgebra of $\mathrm{Mat}_n(\mathbb{Q})$; and, of course, this implies that there are only countably many possibilities for $\mathrm{QE}(A)$. A linear transformation $\varphi \in \mathrm{Mat}_n(\mathbb{Q})$ is said to be a *quasi-automorphism* of A iff φ is a unit of the $\mathbb{Q}$-algebra $\mathrm{QE}(A)$; and the group of quasi-automorphisms of A is denoted by $\mathrm{QAut}(A)$.

Lemma 3.8 (Thomas [31]). *If $A \in R(\mathbb{Q}^n)$, then* $\mathrm{QAut}(A)$ *is the setwise stabilizer of* $[A]$ *in* $\mathrm{GL}_n(\mathbb{Q})$.

In particular, there are only countably many possibilities for the setwise stabilizer of $[A]$ in $\mathrm{GL}_n(\mathbb{Q})$. Hence, arguing as above, we can suppose that there exists a Borel

subset $X_0 \subseteq X$ with $\mu(X_0) = 1$ and a fixed subgroup $L \leqslant \mathrm{GL}_n(\mathbb{Q})$ such that L is the setwise stabilizer of $[A_x]$ for all $x \in X_0$; and this implies that the quotient group $H = N_{\mathrm{GL}_n(\mathbb{Q})}(L)/L$ acts freely on the corresponding set $Y = \{[A] \mid \mathrm{QAut}(A) = L\}$ of $\approx_n$-classes.

Lemma 3.9 (Thomas [31]). *There is an algebraic $\mathbb{Q}$-group G with* $\dim G < m^2 - 1$ *such that $H \leqslant G(\mathbb{Q})$.*

Consequently, we are now positioned to apply Theorem 3.4 … except for one last complication. Unfortunately, the equivalence relation $\approx_n$ is not smooth and this means that Y is not a standard Borel space. However, this turns out not to be a serious difficulty. As shown in Thomas [31], the equivalence relation $\approx_n$ is hyperfinite (which is only slightly more complicated than smooth) and Theorem 3.4 is easily extended to cover induced free actions on quotients of standard Borel spaces by hyperfinite equivalence relations.

Remark 3.10. The above argument does not go through in the case when $n = 1$ because of the failure of Zimmer's cocycle superrigidity theorem for the low rank Lie group $\mathrm{SL}_2(\mathbb{R})$. However, as we mentioned earlier, this case had already been dealt with by Hjorth [13], who gave a completely elementary proof that $(\cong_1) <_B (\cong_2)$, based upon the fact that $\mathrm{GL}_1(\mathbb{Q}) = \mathbb{Q}^*$ is amenable and $\mathrm{GL}_2(\mathbb{Q})$ is nonamenable.

In the remainder of this section, we shall sketch the proof of Theorem 1.14. This involves trying to understand the orbit equivalence relation $\cong_2^p$ of the classical group $\mathrm{GL}_2(\mathbb{Q})$ on the highly non-classical space $R^p(\mathbb{Q}^2)$ of p-local torsion-free abelian groups of rank 2. Fortunately, using the invariants of Kurosh–Malcev [22], [25], it is possible to replace $R^p(\mathbb{Q}^2)$ by a much more intelligible space.

Definition 3.11. For each prime p, let E_p be the orbit equivalence relation arising from the natural action of $\mathrm{GL}_2(\mathbb{Q})$ on the projective line $\mathbb{Q}_p \cup \{\infty\}$ over the field of p-adic numbers.

Theorem 3.12 (Thomas [31]). $(\cong_2^p) \sim_B (E_p)$.

Thus Theorem 1.14 is an immediate consequence of the following result.

Theorem 3.13 (Hjorth–Thomas [16]). *If $p \neq q$ are distinct primes, then the orbit equivalence relations E_p, E_q of* $\mathrm{GL}_2(\mathbb{Q})$ *on the projective lines $\mathbb{Q}_p \cup \{\infty\}$, $\mathbb{Q}_q \cup \{\infty\}$ are incomparable with respect to Borel reducibility.*

Sketch proof of Theorem 3.12. Following Kurosh–Malcev [22], [25], we shall describe how to assign points $V_A \in \mathbb{Q}_p \cup \{\infty\}$ to the p-local groups

$$\{A \in R^p(\mathbb{Q}^2) \mid A \ncong \mathbb{Q} \oplus \mathbb{Q},\ \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}\}$$

such that:

- $A \cong B$ iff the corresponding points V_A, V_B lie in the same $\mathrm{GL}_2(\mathbb{Q})$-orbit;

- for each point $V \in \mathbb{Q}_p \cup \{\infty\}$, there exists a corresponding group A such that $V_A = V$.

The result then follows easily from the fact that there are only countably many groups $A \in R^p(\mathbb{Q}^2)$ such that $A \cong \mathbb{Q} \oplus \mathbb{Q}, \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}$.

It is first necessary to discuss the p-adic completion $\widehat{A}$ of each p-local group $A \in R^p(\mathbb{Q}^2)$, which is defined as follows. For the remainder of this proof, we shall regard $\mathbb{Q}^2$ as an additive subgroup of the 2-dimensional vector space $\mathbb{Q}_p^2$ over the field of p-adic numbers; and we shall regard $\mathrm{GL}_2(\mathbb{Q})$ as a subgroup of $\mathrm{GL}_2(\mathbb{Q}_p)$. For each $A \in R^p(\mathbb{Q}^2)$, let $\widehat{A} = \mathbb{Z}_p \otimes A$; i.e. $\widehat{A}$ is the subgroup of $\mathbb{Q}_p^2$ consisting of all finite sums

$$\gamma_1 a_1 + \gamma_2 a_2 + \cdots + \gamma_t a_t,$$

where $\gamma_i \in \mathbb{Z}_p$ and $a_i \in A$ for $1 \leq i \leq t$. Then, while A usually has a very complex structure, $\widehat{A}$ always decomposes into a direct sum of copies of $\mathbb{Z}_p$ and $\mathbb{Q}_p$. In fact, assuming that $A \ncong \mathbb{Q} \oplus \mathbb{Q}, \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}$, there exist elements $\mathbf{v}_A, \mathbf{w}_A \in \widehat{A}$ such that

$$\widehat{A} = \mathbb{Q}_p \mathbf{v}_A \oplus \mathbb{Z}_p \mathbf{w}_A.$$

(See Fuchs [8].) Let $V_A = \mathbb{Q}_p \mathbf{v}_A$. If $A \cong B$, then there exists $\varphi \in \mathrm{GL}_2(\mathbb{Q})$ such that $\varphi[A] = B$. This implies that $\varphi[\widehat{A}] = \widehat{B}$ and it follows easily that $\varphi[V_A] = V_B$. Conversely, suppose that there exists $\varphi \in \mathrm{GL}_2(\mathbb{Q})$ such that $\varphi[V_A] = V_B$. Since the nontrivial proper $\mathbb{Z}_p$-submodules of $\mathbb{Q}_p$ are precisely $\{p^\ell \mathbb{Z}_p \mid \ell \in \mathbb{Z}\}$, after composing φ with a suitable transformation $v \mapsto p^\ell v$ if necessary, we can suppose that $\varphi[\widehat{A}] = \widehat{B}$. Since $\widehat{A} \cap \mathbb{Q}^2 = A$ and $\widehat{B} \cap \mathbb{Q}^2 = B$, it follows that $\varphi[A] = B$. Thus the $\mathrm{GL}_2(\mathbb{Q})$-orbit of the point $V_A \in \mathbb{Q}_p \cup \{\infty\}$ is a complete invariant for those $A \in R^p(\mathbb{Q}^2)$ such that $A \ncong \mathbb{Q} \oplus \mathbb{Q}, \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}$. □

Remark 3.14. It is now easy to prove that for each prime $p \in \mathbb{P}$, there exists a nonatomic $\mathrm{SL}_2(\mathbb{Z})$-invariant ergodic probability measure μ_p on $R(\mathbb{Q}^2)$ which concentrates on the Borel subspace $R^p(\mathbb{Q}^2)$ of p-local groups. We have just seen how to assign a corresponding point $V_A \in \mathbb{Q}_p \cup \{\infty\}$ to each $A \in R^p(\mathbb{Q}^2)$ such that $A \ncong \mathbb{Q} \oplus \mathbb{Q}, \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}$. Conversely, for each point $V \in \mathbb{Q}_p \cup \{\infty\}$, there exists a corresponding group A such that $V_A = V$. In fact, there exist countably many such groups. However, if we restrict our attention to the $\mathrm{SL}_2(\mathbb{Z})$-invariant Borel subset $X(\mathbb{Q}^2)$ consisting of those $A \in R^p(\mathbb{Q}^2)$ such that

(i) $A \ncong \mathbb{Q} \oplus \mathbb{Q}, \mathbb{Z}_{(p)} \oplus \mathbb{Z}_{(p)}$,

(ii) $\mathbb{Z}_{(p)}^2 \leqslant A$ and $\mathbb{Z}_{(p)}^2 \not\leqslant pA$,

then we obtain a one-to-one correspondence. In summary, the map

$$\begin{aligned} X(\mathbb{Q}^m) &\to \mathbb{Q}_p \cup \{\infty\}, \\ A &\mapsto V_A, \end{aligned}$$

is a Borel bijection satisfying $\varphi[V_A] = V_{\varphi[A]}$ for all $\varphi \in \mathrm{SL}_2(\mathbb{Z})$ and $A \in X(\mathbb{Q}^2)$. Hence the result follows from the observation that there exists a nonatomic $\mathrm{SL}_2(\mathbb{Z})$-invariant ergodic probability measure ν_p on $\mathbb{Q}_p \cup \{\infty\}$. To see this, recall that the compact group $K = \mathrm{SL}_2(\mathbb{Z}_p)$ acts transitively on $\mathbb{Q}_p \cup \{\infty\}$. Hence, letting L be the stabilizer in K of some point of $\mathbb{Q}_p \cup \{\infty\}$, we can identify the K-spaces $\mathbb{Q}_p \cup \{\infty\}$ and K/L. Let ν_p be the Haar probability measure on K/L. Since $\mathrm{SL}_2(\mathbb{Z})$ is a dense subgroup of K, it follows that ν_p is the unique $\mathrm{SL}_2(\mathbb{Z})$-invariant probability measure on K/L and hence $\mathrm{SL}_2(\mathbb{Z})$ acts ergodically on $(K/L, \nu_p)$.

The above argument easily generalizes to show that for all $m \geq 2$, there exists a nonatomic $\mathrm{SL}_m(\mathbb{Z})$-invariant ergodic probability measure μ_p on $R(\mathbb{Q}^m)$ which concentrates on the Borel subspace consisting of those $A \in R^p(\mathbb{Q}^m)$ such that $\dim A/pA = 1$. (For example, see Thomas [31].)

Finally we shall sketch the proof of Theorem 3.13. Recall that if $S = \{p_1, \ldots, p_t\}$ is a nonempty finite set of primes, then

$$\Gamma_S = \mathrm{SL}_2(\mathbb{Z}[1/p_1, \ldots, 1/p_t]).$$

Also let $\Gamma_\emptyset = \mathrm{SL}_2(\mathbb{Z})$. As we shall see, Theorem 3.13 is an easy consequence of the following variant of Theorem 2.16, together with a crucial result of Hjorth [16].

Theorem 3.15 (Thomas [34]). *Suppose that $p \neq q$ are distinct primes and that S is a (possibly empty) finite set of primes. Let*

- *E_1 be the orbit equivalence relation induced by the action of $\mathrm{SL}_2(\mathbb{Z}[1/q])$ on $\mathbb{Q}_p \cup \{\infty\}$, and*
- *E_2 be the orbit equivalence relation induced by the action of Γ_S on $\mathbb{Q}_q \cup \{\infty\}$.*

If $f : \mathbb{Q}_p \cup \{\infty\} \to \mathbb{Q}_q \cup \{\infty\}$ is a Borel homomorphism from E_1 to E_2, then there exists a μ_p-measure 1 subset which is mapped to a single E_2-class.

Remark 3.16. The basic theme of Borel superigidity theory is that, under suitably strong hypotheses, every nontrivial Borel homomorphism is a "slight perturbation" of a virtual homomorphism of the corresponding measure-preserving permutation groups. In the statement of Theorem 3.15, the group $\mathrm{SL}_2(\mathbb{Z}[1/q])$ was chosen because its actions on $\mathbb{Q}_p \cup \{\infty\}$ and $\mathbb{Q}_q \cup \{\infty\}$ are extremely incompatible; namely, while $\mathrm{SL}_2(\mathbb{Z}[1/q])$ preserves the p-adic probability measure on $\mathbb{Q}_p \cup \{\infty\}$, there are no $\mathrm{SL}_2(\mathbb{Z}[1/q])$-invariant probability measures on $\mathbb{Q}_q \cup \{\infty\}$.

Sketch proof of Theorem 3.13. Suppose that $f : \mathbb{Q}_p \cup \{\infty\} \to \mathbb{Q}_q \cup \{\infty\}$ is a Borel reduction between the orbit equivalence relations induced by the $\mathrm{GL}_2(\mathbb{Q})$-actions. Then we can regard f as a countable-to-one Borel homomorphism between the $\mathrm{SL}_2(\mathbb{Z}[1/q])$-action on $\mathbb{Q}_p \cup \{\infty\}$ and the $\mathrm{GL}_2(\mathbb{Q})$-action on $\mathbb{Q}_q \cup \{\infty\}$. Using a suitable Cocycle Reduction Theorem of Hjorth [16], we can "adjust" f to obtain a

countable-to-one Borel homomorphism $f' \colon \mathbb{Q}_p \cup \{\infty\} \to \mathbb{Q}_q \cup \{\infty\}$ between the orbit equivalence relation induced by the $\mathrm{SL}_2(\mathbb{Z}[1/q])$-action on $\mathbb{Q}_p \cup \{\infty\}$ and the orbit equivalence relation induced by the Γ_S-action on $\mathbb{Q}_q \cup \{\infty\}$ for some finite set of primes S, which contradicts Theorem 3.15. □

In view of Remark 1.17, it would be interesting to know whether $\cong_2^p$ is an immediate successor of $\cong_1$ with respect to $\leq_B$. Equivalently:

Question 3.17. Let E_p be the orbit equivalence relation arising from the action of $\mathrm{GL}_2(\mathbb{Q})$ on the projective line $\mathbb{Q}_p \cup \{\infty\}$ over the field of p-adic numbers. Does there exist a (countable) Borel equivalence relation E such that $E_0 <_B E <_B E_p$?

References

[1] Adams, S., Containment does not imply Borel reducibility. In *Set Theory: The Hajnal Conference* (ed. by S. Thomas), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 58, Amer. Math. Soc., Providence, RI, 2002, 1–23.

[2] Adams, S. R., Kechris, A. S., Linear algebraic groups and countable Borel equivalence relations. *J. Amer. Math. Soc.* **13** (2000), 909–943.

[3] Baer, R., Abelian groups without elements of finite order. *Duke Math. J.* **3** (1937), 68–122.

[4] Champetier, C., L'espace des groupes de type fini. *Topology* **39** (2000), 657–680.

[5] Dougherty, R., Jackson, S., Kechris, A. S., The structure of hyperfinite Borel equivalence relations. *Trans. Amer. Math. Soc.* **341** (1994), 193–225.

[6] Feldman, J., Moore, C. C., Ergodic equivalence relations, cohomology and von Neumann algebras. I, *Trans. Amer. Math. Soc.* **234** (1977), 289–324.

[7] Friedman, H., Stanley, L., A Borel reducibility theory for classes of countable structures. *J. Symbolic Logic* **54** (1989), 894–914.

[8] Fuchs, L., *Infinite Abelian Groups*. Vol. II, Pure and Applied Mathematics 36, Academic Press, New York, London 1973.

[9] Furman, A., Orbit equivalence rigidity. *Ann. Math.* **150** (1999), 1083–1108.

[10] Gefter, S. L., Golodets, V. Ya., Fundamental groups for ergodic actions and actions with unit fundamental groups. *Publ. Res. Inst. Math. Sci.* **24** (1988), 821–847.

[11] Gaboriau, D., Popa, S., An uncountable family of nonorbit equivalent actions of $\mathbb{F}_n$. *J. Amer. Math. Soc.* **18** (2005), 547–559.

[12] Harrington, L., Kechris, A. S., Louveau, A., A Glimm-Effros dichotomy for Borel equivalence relations. *J. Amer. Math. Soc.* **3** (1990), 903–927.

[13] Hjorth, G., Around nonclassifiability for countable torsion-free abelian groups. In *Abelian groups and modules* (Dublin, 1998), Trends Math., Birkhäuser, Basel 1999, 269–292.

[14] Hjorth, G., Kechris, A. S., Borel equivalence relations and classification of countable models. *Ann. Pure Appl. Logic* **82** (1996), 221–272.

[15] Hjorth, G., Kechris, A. S., Rigidity theorems for actions of product groups and countable Borel equivalence relations. *Mem. Amer. Math. Soc.* **177** (2005), no. 833.

[16] Hjorth, G., Thomas, S., The classification problem for p-local torsion-free abelian groups of rank two. Preprint, 2004.

[17] Jackson, S., Kechris, A. S., Louveau, A., Countable Borel equivalence relations. *J. Math. Logic* **2** (2002), 1–80.

[18] Jónsson, B., On direct decompositions of torsion-free abelian groups. *Math. Scand.* **7** (1959), 361–371.

[19] Kechris, A. S., Countable sections for locally compact group actions. *Ergodic Theory Dynam. Systems* **12** (1992), 283–295.

[20] Kechris, A. S., *Classical Descriptive Set Theory*. Grad. Texts in Math. 156, Springer-Verlag, New York 1995.

[21] Kuratowski, K., *Topology*. Vol. I, Academic Press, New York, London 1966.

[22] Kurosh, A. G., Primitive torsionsfreie abelsche Gruppen vom endlichen Range. *Ann. Math.* **38** (1937), 175–203.

[23] Lubotzky, A., *Discrete Groups, Expanding Graphs and Invariant Measures*. Progr. Math. 125, Birkhäuser, Basel 1994.

[24] Lyndon, R. C., Schupp, P. E., *Combinatorial Group Theory*. Ergeb. Math. Grenzgeb. 89, Springer-Verlag, Berlin, New York 1977.

[25] Malcev, A. I., Torsion-free abelian groups of finite rank (Russian). *Mat. Sb.* **4** (1938), 45–68.

[26] Margulis, G. A., Tomanov, G. M., Measure rigidity for almost linear groups and its applications. *J. Anal. Math.* **69** (1996), 25–54.

[27] Ornstein, D., Weiss, B., Ergodic theory of amenable group actions. I. The Rohlin lemma. *Bull. Amer. Math. Soc.* **2** (1980), 161–164.

[28] Platonov, V., Rapinchuk, A., *Algebraic Groups and Number Theory*. Pure Appl. Math., Academic Press, Boston, MA, 1994.

[29] Ratner, M., On Raghunathan's measure conjecture. *Ann. Math.* **134** (1991), 545–607.

[30] Solovay, R. M., A model of set theory in which every set of reals is Lebesgue measurable. *Ann. Math.* **92** (1970), 1–56.

[31] Thomas, S., The classification problem for torsion-free abelian groups of finite rank. *J. Amer. Math. Soc.* **16** (2003), 233–258.

[32] Thomas, S., Superrigidity and countable Borel equivalence relations. *Ann. Pure Appl. Logic* **120** (2003), 237–262.

[33] Thomas, S., The classification problem for p-local torsion-free abelian groups of finite rank. Preprint, 2002.

[34] Thomas, S., Property (τ) and countable Borel equivalence relations. Preprint, 2004.

[35] Thomas, S., The classification problem for torsion-free abelian groups of finite rank II. In preparation.

[36] Thomas, S., Velickovic, B., On the complexity of the isomorphism relation for finitely generated groups. *J. Algebra* **217** (1999), 352–373.

[37] Thomas, S., Velickovic, B., On the complexity of the isomorphism relation for fields of finite transcendence degree. *J. Pure Appl. Algebra.* **159** (2001), 347–363.

[38] Zimmer, R. J., *Ergodic Theory and Semisimple Groups*. Monogr. Math. 81, Birkhäuser, Basel 1984.

[39] Zimmer, R. J., Superrigidity, Ratner's theorem, and fundamental groups. *Israel J. Math.* **74** (1991), 199–207.

Mathematics Department, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854-8019, U.S.A.

E-mail: sthomas@math.rutgers.edu

Quiver algebras, weighted projective lines, and the Deligne–Simpson problem

William Crawley-Boevey

Abstract. We describe recent work on preprojective algebras and moduli spaces of their representations. We give an analogue of Kac's Theorem, characterizing the dimension types of indecomposable coherent sheaves over weighted projective lines in terms of loop algebras of Kac–Moody Lie algebras, and explain how it is proved using Hall algebras. We discuss applications to the problem of describing the possible conjugacy classes of sums and products of matrices in known conjugacy classes.

Mathematics Subject Classification (2000). Primary 16G20, 14H60, 15A24.

Keywords. Quiver, Kac–Moody Lie algebra, preprojective algebra, weighted projective line, parabolic bundle, loop algebra, Hall algebra, Deligne–Simpson problem.

Introduction

Preprojective algebras were introduced by Gelfand and Ponomarev, and in a deformed version by Crawley-Boevey and Holland. They arose in the theory of representations of quivers, but have interesting links with Kleinian singularities, Kac–Moody Lie algebras and noncommutative symplectic geometry. In the first part of this article, §1, we survey some of the results we have obtained in the last ten years concerning these algebras, and moduli spaces of their representations.

The Deligne–Simpson problem asks about the existence of matrices in given conjugacy classes, with product the identity, and no common invariant subspace. Some time ago it became clear that our work on preprojective algebras solves an additive analogue. To solve the original problem, one needs to pass to a new setup, in which representations of quivers are replaced by coherent sheaves on weighted projective lines (or parabolic bundles), and representations of the preprojective algebra are replaced by logarithmic connections. We discuss all this in §4.

A key ingredient in the theory of preprojective algebras is Kac's Theorem, describing the possible dimension vectors of indecomposable representations of quivers. In the new setup, one needs an analogue of Kac's Theorem for weighted projective lines. We discuss it in §2, and outline a proof via Hall algebras in §3.

In the rest of this introduction we recall some basic facts about representations of quivers. A *quiver* Q, or more precisely (I, Q, h, t), consists of finite sets I and Q

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

of vertices and arrows, and maps $h, t\colon Q \to I$, assigning to each arrow its head and tail vertices. We fix a base field K, algebraically closed unless otherwise indicated. By a *representation* X of Q, one means the assignment of a vector space X_v for each vertex v, and a linear map $X_{t(a)} \to X_{h(a)}$ for each arrow a. There are natural notions of homomorphisms between representations, sub-representations, etc.

The *path algebra* KQ has basis the paths $a_1a_2\dots a_n$ in Q of length $n \geq 1$, with $t(a_i) = h(a_{i+1})$ for all i, and a *trivial path* of length 0 for each vertex v. It is an associative algebra, with the product of two paths given by their concatenation, if this makes sense, and otherwise zero, and the sum of the trivial paths is a multiplicative identity. The category of representations of Q is equivalent to the category of left KQ-modules, so one can use homological algebra, composition series, the Krull–Remak–Schmidt Theorem, and so on.

We now assume for simplicity that Q has no oriented cycles, in which case KQ is finite dimensional, although many results hold without this restriction.

Let $\mathfrak{g}$ be the Kac–Moody Lie algebra given by the symmetric generalized Cartan matrix $(a_{uv})_{u,v\in I}$ which has diagonal entries 2, and off-diagonal entries given by minus the number of arrows in Q between u and v, in either direction. Thus $\mathfrak{g}$ is generated over $\mathbb{C}$ by e_v, f_v, h_v ($v \in I$) with relations

$$\begin{cases} [h_u, h_v] = 0, \quad [e_u, f_v] = \delta_{uv} h_v, \\ [h_u, e_v] = a_{uv} e_v, \quad [h_u, f_v] = -a_{uv} f_v, \\ (\operatorname{ad} e_u)^{1-a_{uv}}(e_v) = 0, \quad (\operatorname{ad} f_u)^{1-a_{uv}}(f_v) = 0 \quad (\text{if } u \neq v), \end{cases} \tag{1}$$

where δ is the Kronecker delta function. The root lattice Γ of $\mathfrak{g}$ is the free additive group on symbols α_v ($v \in I$), it grades $\mathfrak{g}$, with $\deg e_v = \alpha_v$, $\deg f_v = -\alpha_v$ and $\deg h_v = 0$, and the set of roots is $\Delta = \{0 \neq \alpha \in \Gamma \mid \mathfrak{g}_\alpha \neq 0\}$. Recall that there are real roots, obtained from the simple roots α_u by a sequence of reflections $s_v(\alpha) = \alpha - (\alpha, \alpha_v)\alpha_v$, where $(-,-)$ is the symmetric bilinear form on Γ with $(\alpha_u, \alpha_v) = a_{uv}$, and there may also be imaginary roots. Defining $p(\alpha) = 1 - \frac{1}{2}(\alpha, \alpha)$, the real roots have $p(\alpha) = 0$, and the imaginary roots have $p(\alpha) > 0$.

Gabriel's Theorem [19] asserts that a quiver Q has only finitely many indecomposable representations if and only if $\mathfrak{g}$ is of finite type, i.e. the underlying graph of Q is a Dynkin diagram (of type ADE). In this case the map sending a representation to its dimension vector

$$\underline{\dim}\, X = \sum_{v\in I} (\dim X_v)\alpha_v \in \Gamma$$

gives a 1-1 correspondence between indecomposable representations and positive roots. Kac's Theorem [26], [27] extends this to $\mathfrak{g}$ of arbitrary type: the dimension vectors of indecomposable representations are exactly the positive roots, there is a unique indecomposable for each real root, infinitely many for each imaginary root.

1. Preprojective algebras

The *preprojective algebra* of a quiver Q is the algebra

$$\Pi(Q) = K\bar{Q}\Big/\Big(\sum_{a\in Q}(aa^* - a^*a)\Big),$$

where the *double* $\bar{Q}$ of Q is obtained by adjoining a reverse arrow a^* for each arrow $a \in Q$. In the finite type case it is isomorphic, as a KQ-module, to the direct sum of one copy of each indecomposable representation of Q. In general, it is isomorphic to the direct sum of the indecomposable representations of Q that are *preprojective*, meaning that some power of the Coxeter functor, or equivalently of the Auslander–Reiten translation $D\mathrm{Tr}$, sends them to a projective KQ-module.

Preprojective algebras first appeared in unpublished work of I. M. Gelfand and V. A. Ponomarev, in a lecture delivered by A. V. Roiter at the Second International Conference on Representations of Algebras (Ottawa, 1979). See [45] for a discussion about variations on this definition. Note also that work by Riedtmann [40] contains parallel ideas.

Given $\lambda = (\lambda_v)_{v\in I} \in K^I$, Crawley-Boevey and Holland [15] have introduced the *deformed preprojective algebra*, $\Pi^\lambda(Q)$, in which the relation is replaced by

$$\sum_{a\in Q}(aa^* - a^*a) - \lambda, \tag{2}$$

where λ is identified with the corresponding linear combination of trivial paths.

Up to isomorphism, these algebras do not depend on the orientation of Q. They are related to some elementary symplectic geometry. Choosing bases for the vector spaces, representations of Q of dimension vector $\alpha = \sum_v n_v\alpha_v$ are given by elements of the space

$$\mathrm{Rep}(Q,\alpha) = \bigoplus_{a\in Q}\mathrm{Mat}_{n_{h(a)}\times n_{t(a)}}(K),$$

and isomorphism classes correspond to orbits of the group

$$\mathrm{GL}(\alpha) = \prod_{v\in I}\mathrm{GL}_{n_v}(K)$$

acting by conjugation. The space of representations of $\bar{Q}$ can then be identified with a cotangent bundle

$$\mathrm{Rep}(\bar{Q},\alpha) \cong \mathrm{Rep}(Q,\alpha)\times\mathrm{Rep}(Q,\alpha)^* \cong T^*\,\mathrm{Rep}(Q,\alpha).$$

This has a natural symplectic structure, and associated to the action of $\mathrm{GL}(\alpha)$ there is a moment map

$$\mu_\alpha\colon \mathrm{Rep}(\bar{Q},\alpha)\to\mathrm{gl}(\alpha),\quad x\mapsto\Big(\sum_{\substack{a\in Q\\ h(a)=v}}x_ax_{a^*} - \sum_{\substack{a\in Q\\ t(a)=v}}x_{a^*}x_a\Big)_{v\in I}.$$

Identifying λ with a central element of $\mathrm{gl}(\alpha)$, there is a quotient

$$N_Q(\lambda, \alpha) = \mu_\alpha^{-1}(\lambda) /\!\!/ \mathrm{GL}(\alpha),$$

a 'symplectic quotient', or 'Marsden–Weinstein reduction'. Here the double slash denotes the affine variety that classifies closed orbits of $\mathrm{GL}(\alpha)$ on $\mu_\alpha^{-1}(\lambda)$. Now the elements of this fibre are exactly the representations of $\overline{Q}$ satisfying (2), so $N_Q(\lambda, \alpha)$ classifies isomorphism classes of semisimple representations of $\Pi^\lambda(Q)$ of dimension vector α.

Since the trace of $\mu_\alpha(x)$ is always zero, if there is a representation of $\Pi^\lambda(Q)$ of dimension vector $\alpha = \sum n_v \alpha_v$, then $\lambda \cdot \alpha = \sum \lambda_v n_v$ must be zero. If Q is a Dynkin diagram, then $\Pi^\lambda(Q)$ is finite-dimensional, and this argument shows that for generic λ it is even the zero algebra.

The case when Q is an extended Dynkin diagram, or equivalently when $\mathfrak{g}$ is an affine Lie algebra, appeared in work of Kronheimer [29], made more explicit by Cassens and Slodowy [6]. If δ is the minimal positive imaginary root, then $N_Q(0, \delta)$ is the corresponding Kleinian surface singularity, and the spaces $N_Q(\lambda, \delta)$, for suitably varying λ, give its semiuniversal deformation. The key idea of [15] is that the deformed preprojective algebras $\Pi^\lambda(Q)$, for unrestricted λ, give a larger family of deformations of the Kleinian singularity, the general one being noncommutative.

Theorem 1 (Crawley-Boevey and Holland [15]). *Suppose Q is an extended Dynkin diagram, and e is the trivial path corresponding to an extending vertex for Q. Then $\mathcal{O}^\lambda = e\Pi^\lambda(Q)e$ is a noetherian domain of Gelfand–Kirillov dimension 2, Auslander-Gorenstein and Cohen–Macaulay. Moreover, it is commutative if and only if $\lambda \cdot \delta = 0$, and if so, it is isomorphic to the coordinate ring of $N_Q(\lambda, \delta)$.*

For some further work on the $\mathcal{O}^\lambda$, see [1], [4], [24]. Returning to the general case, to decide when $N_Q(\lambda, \alpha)$ is nonempty, one needs to know whether or not there is a representation of $\Pi^\lambda(Q)$ of dimension vector α. It is not hard to show that a representation X of Q is in the image of the projection $\mu_\alpha^{-1}(\lambda) \to \mathrm{Rep}(Q, \alpha)$, so is the restriction of a representation of $\Pi^\lambda(Q)$, if and only if the dimension vector β of each indecomposable direct summand of X satisfies $\lambda \cdot \beta = 0$. With Kac's Theorem, this gives the following.

Theorem 2 ([7]). *The space $N_Q(\lambda, \alpha)$ is nonempty, or equivalently there is a representation of $\Pi^\lambda(Q)$ of dimension vector α, if and only if α can be written as a sum of positive roots $\alpha = \beta + \gamma + \cdots$ with $\lambda \cdot \beta = \lambda \cdot \gamma = \cdots = 0$.*

More difficult is the following.

Theorem 3 ([7]). *There is a simple representation of $\Pi^\lambda(Q)$ of dimension vector α if and only if α is a positive root, $\lambda \cdot \alpha = 0$, and $p(\alpha) > p(\beta) + p(\gamma) + \cdots$ for any nontrivial decomposition of α as a sum of positive roots $\alpha = \beta + \gamma + \cdots$ with $\lambda \cdot \beta = \lambda \cdot \gamma = \cdots = 0$.*

We have also shown [8], [9] that if K has characteristic zero and $N_Q(\lambda, \alpha)$ is nonempty, then it is an irreducible normal variety. Bocklandt and Le Bruyn [2], [30] have obtained further results in this direction. See [14] for more about the link with noncommutative symplectic geometry.

There are more general moduli spaces $N_Q(\lambda, \alpha)_\theta$, depending on suitable stability data $\theta \in \mathbb{Z}^I$. Examples of these are the 'quiver varieties' used by Nakajima to construct integrable representations of Lie algebras and quantum groups, see his ICM talk [36] (and [7]). Note that over $\mathbb{C}$, the moduli spaces are special cases of hyper-Kähler quotients, and by a standard trick of changing the complex structure, $N_Q(0, \alpha)_\theta$ is homeomorphic to $N_Q(\theta, \alpha)$.

We now give an application of these ideas. When working over a finite field, it is natural to consider representations of Q which are *absolutely indecomposable*, meaning that they remain indecomposable over the algebraic closure of the field. Kac showed that up to isomorphism, the number such representations of dimension vector α is polynomial in the size q of the field, of the form $a_\alpha(q)$ for some $a_\alpha \in \mathbb{Z}[t]$. He conjectured that a_α has non-negative coefficients, and that the constant term is the root multiplicity $\dim \mathfrak{g}_\alpha$. In partial answer we have the following.

Theorem 4 (Crawley-Boevey and Van den Bergh [17]). *If $\alpha = \sum_{v \in I} n_v \alpha_v$ is indivisible, meaning that the n_v have no common divisor, then a_α has non-negative coefficients, and the constant term is the root multiplicity* $\dim \mathfrak{g}_\alpha$.

We explain the positivity. Since α is indivisible, one can fix $\lambda \in \mathbb{Z}^I$ with $\lambda \cdot \alpha = 0$ but $\lambda \cdot \beta \neq 0$ for all $0 < \beta < \alpha$. The argument of Theorem 2 shows that the number of points in $N_Q(\lambda, \alpha)$, over a field with q elements and sufficiently large characteristic, is $q^{p(\alpha)} a_\alpha(q)$, and then if $N_Q(\lambda, \alpha)$ had been a projective variety, the Weil conjectures would have given positivity. However, $N_Q(0, \alpha)_\lambda$ is sufficiently close to being projective for the Weil conjectures to apply to it, and by the hyper-Kähler trick, the cohomologies of $N_Q(0, \alpha)_\lambda$ and $N_Q(\lambda, \alpha)$ are isomorphic when the base field is the algebraic closure of a finite field of sufficiently large characteristic. Moreover, it is possible to ensure that this isomorphism is compatible with Frobenius maps, so that $N_Q(\lambda, \alpha)$ is good enough.

2. Weighted projective lines

In this section we give an analogue of Kac's Theorem for weighted projective lines. When studying representations of finite-dimensional associative algebras, quivers tell one about *hereditary* algebras, i.e. those with global dimension ≤ 1. One of the breakthroughs in this area was the discovery by Brenner and Butler [5], [23] of algebras A that are 'tilted' from a hereditary algebra H. In Happel's language [22], there is a derived equivalence $D^b(\operatorname{mod} A) \simeq D^b(\operatorname{mod} H)$, which is useful since in $D^b(\operatorname{mod} H)$ any indecomposable object is a shift of a module.

Geigle and Lenzing [20] realized that there are other algebras, including Ringel's 'canonical algebras' [42], which aren't necessarily tilted from hereditary algebras, but are tilted from suitable hereditary abelian categories. See Reiten's ICM talk [39] for further progress in this direction.

We concentrate on Geigle and Lenzing's categories. A *weighted projective line* $\mathbb{X}$ is specified by giving a collection of distinct points $D = (a_1, \dots, a_k)$ in the projective line $\mathbb{P}^1$ over K, and a *weight sequence* $\boldsymbol{w} = (w_1, \dots, w_k)$, that is, a sequence of positive integers. The category, $\operatorname{coh}\mathbb{X}$, of coherent sheaves on $\mathbb{X}$, can be defined as the quotient of the category of finitely generated $\boldsymbol{L}(\boldsymbol{w})_+$-graded $S(\boldsymbol{w}, D)$-modules by the Serre subcategory of finite length modules. Here $\boldsymbol{L}(\boldsymbol{w})$ is the additive group with generators $\vec{x}_1, \dots, \vec{x}_k, \vec{c}$ and relations $w_1\vec{x}_1 = \cdots = w_k\vec{x}_k = \vec{c}$, partially ordered, with positive cone $\boldsymbol{L}(\boldsymbol{w})_+ = \mathbb{N}\vec{c} + \sum_{i=1}^{k} \mathbb{N}\vec{x}_i$, and

$$S(\boldsymbol{w}, D) = K[u, v, x_1, \dots, x_k]/(x_i^{w_i} - \lambda_i u - \mu_i v),$$

where $a_i = [\lambda_i : \mu_i] \in \mathbb{P}^1$, with grading $\deg u = \deg v = \vec{c}$ and $\deg x_i = \vec{x}_i$. Geigle and Lenzing showed that $\operatorname{coh}\mathbb{X}$ is a hereditary abelian category; the free module gives a structure sheaf $\mathcal{O}$, and shifting the grading gives twists $E(\vec{x})$ for any sheaf E and $\vec{x} \in \boldsymbol{L}(\boldsymbol{w})$; also, every sheaf is the direct sum of a 'torsion-free' sheaf, with a filtration by sheaves of the form $\mathcal{O}(\vec{x})$, and a finite-length 'torsion' sheaf.

Let $Q_{\boldsymbol{w}}$ be the star-shaped quiver whose vertex set I consists of a central vertex $*$, and vertices, denoted ij or i, j, for $1 \le i \le k$, $1 \le j < w_i$, and with arrows $* \leftarrow i1 \leftarrow i2 \leftarrow \cdots$ for all i. The appropriate Lie algebra to consider is the loop algebra $L\mathfrak{g} = \mathfrak{g}[t, t^{-1}]$, where $\mathfrak{g}$ is the Kac–Moody algebra associated $Q_{\boldsymbol{w}}$, or, better, an extension $\mathcal{L}\mathfrak{g}$ with generators e_{vr}, f_{vr}, h_{vr} ($v \in I$, $r \in \mathbb{Z}$) and c subject to the relations

$$\begin{cases} c \text{ central}, \quad [e_{vr}, e_{vs}] = 0, \quad [f_{vr}, f_{vs}] = 0, \\ [h_{ur}, h_{vs}] = r a_{uv}\, \delta_{r+s,0}\, c, \quad [e_{ur}, f_{vs}] = \delta_{uv} \left(h_{v,r+s} + r\, \delta_{r+s,0}\, c\right), \\ [h_{ur}, e_{vs}] = a_{uv} e_{v,r+s}, \quad [h_{ur}, f_{vs}] = -a_{uv} f_{v,r+s}, \\ (\operatorname{ad} e_{u0})^{1-a_{uv}}(e_{vs}) = 0, \quad (\operatorname{ad} f_{u0})^{1-a_{uv}}(f_{vs}) = 0 \quad (\text{if } u \ne v), \end{cases} \tag{3}$$

see [35]. The root lattice for either algebra is $\hat{\Gamma} = \Gamma \oplus \mathbb{Z}\delta$, with $\deg e_v t^r = \deg e_{vr} = \alpha_v + r\delta$, $\deg f_v t^r = \deg f_{vr} = -\alpha_v + r\delta$, $\deg h_v t^r = \deg h_{vr} = r\delta$ and $\deg c = 0$, and the set of roots for either algebra is

$$\hat{\Delta} = \{\alpha + r\delta \mid \alpha \in \Delta, r \in \mathbb{Z}\} \cup \{r\delta \mid 0 \ne r \in Z\}.$$

The real roots are $\alpha + r\delta$ with α real. (If $\mathfrak{g}$ is of finite type, $\mathcal{L}\mathfrak{g}$ is the corresponding affine Lie algebra, and if $\mathfrak{g}$ is of affine type, $\mathcal{L}\mathfrak{g}$ is a toroidal algebra.)

The Grothendieck group $K_0(\operatorname{coh}\mathbb{X})$ was computed by Geigle and Lenzing, and following Schiffmann [46] it can be identified with $\hat{\Gamma}$. Now $K_0(\operatorname{coh}\mathbb{X})$ is partially ordered, with the positive cone being the classes of objects in $\operatorname{coh}\mathbb{X}$, and this gives a partial ordering on $\hat{\Gamma}$.

Theorem 5 ([12]). *If $\mathbb{X}$ is a weighted projective line, there is an indecomposable coherent sheaf on $\mathbb{X}$ of type $\phi \in \hat{\Gamma}$ if and only if ϕ is a positive root. There is a unique indecomposable for a real root, infinitely many for an imaginary root.*

We remark that there is a classification of the indecomposables if $\mathfrak{g}$ is of finite type [20], or affine type [32]. The latter is essentially equivalent to the classification for tubular algebras, see Ringel's ICM talk [41], [42].

Lenzing [31, §4.2] showed that the category of torsion-free sheaves on $\mathbb{X}$ is equivalent to the category of (quasi) parabolic bundles on $\mathbb{P}^1$ of weight type $(D, \boldsymbol{w})$, that is, vector bundles $\pi : E \to \mathbb{P}^1$ equipped with a flag of vector subspaces

$$\pi^{-1}(a_i) \supseteq E_{i1} \supseteq \cdots \supseteq E_{i,w_i-1}$$

for each i. This equivalence is not unique, but it can be chosen so that if E is a parabolic bundle, then $[E] = \underline{\dim}\, E + (\deg E)\delta$, where the dimension vector is

$$\underline{\dim}\, E = n_*\alpha_* + \sum_{i=1}^{k} \sum_{j=1}^{w_i-1} n_{ij}\alpha_{ij} \in \Gamma,$$

with $n_* = \operatorname{rank} E$ and $n_{ij} = \dim E_{ij}$. This is necessarily *strict*, meaning that $n_* \geq n_{i1} \geq n_{i2} \geq \cdots \geq n_{i,w_i-1} \geq 0$. We can now restate Theorem 5 as follows (see [11]). For each $d \in \mathbb{Z}$ there is an indecomposable parabolic bundle of dimension vector α and degree d if and only if α is a strict root for $\mathfrak{g}$. There is a unique indecomposable for a real root, infinitely many for an imaginary root.

3. Hall algebras

In this section we explain the proof of Theorem 5. Let $\mathcal{C}$ be an abelian category that is *finitary*, meaning that its Hom and Ext spaces are finite sets. The *Hall algebra* of $\mathcal{C}$, over a commutative ring Λ, is the free Λ-module

$$H_\Lambda(\mathcal{C}) = \bigoplus_{Z \in \operatorname{iso}\mathcal{C}} \Lambda u_Z,$$

with basis the symbols u_Z, where Z runs through $\operatorname{iso}\mathcal{C}$, a set of representatives of the isomorphism classes of $\mathcal{C}$. It is an associative algebra with product

$$u_X u_Y = \sum_{Z \in \operatorname{iso}\mathcal{C}} F_{XY}^Z u_Z,$$

where F_{XY}^Z is the number of subobjects Z' of Z with $Z' \cong Y$ and $Z/Z' \cong X$. In case $\mathcal{C}$ is the category of finite abelian groups, or finite abelian p-groups, this notion is due to Steinitz [47] and Hall [21]. The current interest in Hall algebras stems from the discovery by Ringel [44] of a relationship between quantum groups and Hall algebras

for categories of representations of finite-dimensional hereditary algebras over finite fields. This quickly influenced the development of canonical bases, see Lusztig's ICM talk [34].

How to recover the underlying Lie algebra? Ringel [43] realized, for finite type hereditary algebras over a finite field K, that if one uses a ring Λ in which $|K| = 1$, and ind $\mathcal{C}$ is the set of indecomposables in iso $\mathcal{C}$, then the u_Z with $Z \in \operatorname{ind} \mathcal{C}$ generate a Lie subalgebra of $H_\Lambda(\mathcal{C})$ with bracket

$$[u_X, u_Y] = \sum_Z (F^Z_{XY} - F^Z_{YX}) u_Z.$$

To get at something resembling a semisimple Lie algebra, not just its positive part, there is a construction of Peng and Xiao [25], [38], in which one starts not with an abelian category, but with a triangulated K-category that is 2-*periodic*, meaning that the shift functor T satisfies $T^2 = 1$.

Let $\mathbb{X}$ be a weighted projective line over a finite field K, whose marked points are all defined over K. The category $\operatorname{coh} \mathbb{X}$ is still defined and well-behaved, and Schiffmann [46] has studied its Hall algebra. Applying the construction of Peng and Xiao to the orbit category $\mathcal{R}_\mathbb{X} = D^b(\operatorname{coh} \mathbb{X})/(T^2)$, one obtains a Lie algebra with triangular decomposition

$$L_\Lambda(\mathcal{R}_\mathbb{X}) = \Big(\bigoplus_{X \in \operatorname{ind} \operatorname{coh} \mathbb{X}} \Lambda u_X \Big) \oplus (\Lambda \otimes_\mathbb{Z} \hat{\Gamma}) \oplus \Big(\bigoplus_{X \in \operatorname{ind} \operatorname{coh} \mathbb{X}} \Lambda u_{TX} \Big),$$

where Λ is still a commutative ring in which $|K| = 1$. We have the following result.

Theorem 6 ([12]). *$L_\Lambda(\mathcal{R}_\mathbb{X})$ contains elements e_{vr}, f_{vr}, h_{vr} ($v \in I$, $r \in \mathbb{Z}$) and c satisfying the relations* (3) *for $\mathcal{L}\mathfrak{g}$.*

The elements are explicitly given: $c = -1 \otimes \delta$, $e_{*,r} = u_{\mathcal{O}(r\vec{c})}$, $f_{*,r} = -u_{T\mathcal{O}(-r\vec{c})}$, and the $e_{ij,r}$ and $f_{ij,r}$ are all of the form u_X or $-u_{TX}$ for suitable indecomposable torsion sheaves X. See also [33], where elliptic Lie algebra generators are found in $L_\Lambda(\mathcal{R}_\mathbb{X})$ for $\mathfrak{g}$ of affine type.

Concerning the proof of Theorem 5, the main problem is to show that the number of indecomposables of type $\phi = \alpha + r\delta$ is the same as the number of type $s_v(\alpha) + r\delta$. By arguments already used in the proof of Kac's Theorem, one reduces to counting numbers of indecomposables for weighted projective lines over finite fields, so dimensions of root spaces of $L_\Lambda(\mathcal{R}_\mathbb{X})$. A standard argument in the theory of complex Lie algebras, using sl_2-triples (e, f, h), shows that the root multiplicities for roots related by a reflection are equal. Now Theorem 6 provides such triples, and although the argument uses the fact that the base field has characteristic zero, for example it involves the operator $\exp(\operatorname{ad} e)$ with $\operatorname{ad} e$ acting locally nilpotently, it works if Λ is a field of sufficiently large characteristic, and this can be arranged by taking the finite field K to be sufficiently large.

4. The Deligne–Simpson problem

Given invertible matrices in known conjugacy (i.e. similarity) classes, what can one say about the conjugacy class of their product? More symmetrically, given conjugacy classes $C_1, \dots, C_k$ in $\mathrm{GL}_n(\mathbb{C})$, is there a solution to the equation

$$A_1 A_2 \dots A_k = 1 \tag{4}$$

with $A_i \in C_i$? The additive analogue asks for a solution to the equation

$$A_1 + A_2 + \dots + A_k = 0, \tag{5}$$

where the conjugacy classes may now be in $\mathrm{gl}_n(\mathbb{C})$. In full generality these problems seem to be open, but there are partial results, see for example [37]. The former arises when studying linear ODEs

$$\frac{d^n f}{dz^n} + c_1(z)\frac{d^{n-1} f}{dz^{n-1}} + \dots + c_n(z) f = 0 \tag{6}$$

whose coefficients are rational functions of z. If $D = \{a_1, \dots, a_k\}$ is the set of singular points of the coefficients in $\mathbb{P}^1$, the monodromy of (6) is a representation in $\mathrm{GL}_n(\mathbb{C})$ of the fundamental group of the punctured Riemann sphere $\mathbb{P}^1 \setminus D$, and the presentation of this group as $\langle g_1, \dots, g_k \mid g_1 g_2 \dots g_k = 1 \rangle$, where g_i is a suitable loop around a_i, shows how equation (4) arises.

To fix the conjugacy classes we choose a weight sequence $\boldsymbol{w} = (w_1, \dots, w_k)$, and a collection of complex numbers $\xi = (\xi_{ij})$ $(1 \le i \le k,\ 1 \le j \le w_i)$ with $(A_i - \xi_{i1}1)(A_i - \xi_{i2}1) \dots (A_i - \xi_{i,w_i}1) = 0$ for $A_i \in C_i$. Clearly, if one wishes one can take w_i to be the degree of the minimal polynomial of A_i, and $\xi_{i1}, \dots, \xi_{i,w_i}$ to be its roots. Let $Q_{\boldsymbol{w}}$ be the quiver associated to $\boldsymbol{w}$ as in §2, let I be its vertex set, let $\mathfrak{g}$ be the corresponding Kac–Moody Lie algebra, and let Γ be its root lattice. The C_i determine an element $\alpha = \sum_v n_v \alpha_v \in \Gamma$, with $n_* = n$ and

$$n_{ij} = \mathrm{rank}(A_i - \xi_{i1}1)(A_i - \xi_{i2}1) \dots (A_i - \xi_{ij}1)$$

for $A_i \in C_i$, and conversely $\boldsymbol{w}$, ξ, and α determine the C_i. We define

$$\xi^{[\beta]} = \prod_{i=1}^{k} \prod_{j=1}^{w_i} \xi_{ij}^{m_{i,j-1} - m_{ij}}, \quad \xi * [\beta] = \sum_{i=1}^{k} \sum_{j=1}^{w_i} \xi_{ij}(m_{i,j-1} - m_{ij}),$$

for $\beta = \sum m_v \alpha_v$, with the convention that $m_{i0} = m_*$ and $m_{i,w_i} = 0$. Theorem 2, applied to a deformed preprojective algebra $\Pi^\lambda(Q_{\boldsymbol{w}})$, gives the following.

Theorem 7 ([11]). *There is a solution to $A_1 + \dots + A_k = 0$ with A_i in the closure $\overline{C_i}$ of C_i if and only if α can be written as a sum of positive roots $\alpha = \beta + \gamma + \dots$ with $\xi * [\beta] = \xi * [\gamma] = \dots = 0$.*

A solution to equation (4) or (5) is *irreducible* if the A_i have no common invariant subspace. Theorem 3 gives the following.

Theorem 8 ([10]). *There is an irreducible solution to $A_1+\cdots+A_k=0$ with $A_i \in C_i$ if and only if α is a positive root, $\xi*[\alpha]=0$, and $p(\alpha) > p(\beta)+p(\gamma)+\cdots$ for any nontrivial decomposition of α as a sum of positive roots $\alpha=\beta+\gamma+\cdots$ with $\xi*[\beta]=\xi*[\gamma]=\cdots=0$.*

What about the multiplicative equation? By the Riemann–Hilbert correspondence, any solution arises as the monodromy of the differential equation given by a logarithmic connection on a vector bundle for $\mathbb{P}^1$. (Note that a Fuchsian ODE (6), as hoped for in Hilbert's 21st problem in his 1900 ICM talk, will not suffice, nor will a Fuchsian system of linear differential equations, or equivalently a logarithmic connection on a trivial vector bundle, as discussed by Bolibruch in his ICM talk [3].) Now a theorem of Weil [49] asserts that a vector bundle on a compact Riemann surface has a holomorphic connection if and only if its indecomposable direct summands have degree 0. There is an analogous theorem for parabolic bundles and compatible logarithmic connections, see [11], and, using it, Theorem 5 implies the following.

Theorem 9 ([11]). *There is a solution to $A_1 \ldots A_k = 1$ with $A_i \in \overline{C_i}$ if and only if α can be written as a sum of positive roots $\alpha=\beta+\gamma+\cdots$ with $\xi^{[\beta]}=\xi^{[\gamma]}=\cdots=1$.*

The *Deligne–Simpson problem*, see [28], asks when there is an irreducible solution to (4) with $A_i \in C_i$. By considering multiplicative analogues of preprojective algebras, Crawley-Boevey and Shaw deduce the following from Theorem 9.

Theorem 10 (Crawley-Boevey and Shaw [16]). *For there to be an irreducible solution to $A_1 \ldots A_k = 1$ with $A_i \in C_i$ it is sufficient that α be a positive root, $\xi^{[\alpha]}=1$, and $p(\alpha) > p(\beta)+p(\gamma)+\cdots$ for any nontrivial decomposition of α as a sum of positive roots $\alpha=\beta+\gamma+\cdots$ with $\xi^{[\beta]}=\xi^{[\gamma]}=\cdots=1$.*

The condition in the theorem has now also been shown to be necessary [13]. For some recent work related to multiplicative preprojective algebras, see [18] and [48].

References

[1] Baranovsky, V., Ginzburg, V., Kuznetsov, A., Quiver varieties and a noncommutative $\mathbb{P}^2$. *Compositio Math.* **134** (2002), 283–318.

[2] Bocklandt, R., Le Bruyn, L., Necklace Lie algebras and noncommutative symplectic geometry. *Math. Z.* **240** (2002), 141–167.

[3] Bolibruch, A. A., The Riemann-Hilbert problem and Fuchsian differential equations on the Riemann sphere. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 1159–1168.

[4] Boyarchenko, M., Quantization of minimal resolutions of Kleinian singularities. Preprint; math.RT/0505165.

[5] Brenner, S., Butler, M. C. R., Generalizations of the Bernstein-Gelfand-Ponomarev reflection functors. In *Representation Theory II* (Ottawa, 1979), ed. by V. Dlab and P. Gabriel, Lecture Notes in Math. 832, Springer-Verlag, Berlin 1980, 103–169.

[6] Cassens, H., Slodowy, P., On Kleinian singularities and quivers. In *Singularities*. The Brieskorn anniversary volume (Oberwolfach, 1996), ed. by V. I. Arnold et al., Progr. Math. 162, Birkhäuser, Basel 1998, 263–288.

[7] Crawley-Boevey, W., Geometry of the moment map for representations of quivers. *Compositio Math.* **126** (2001), 257–293.

[8] —, Decomposition of Marsden-Weinstein reductions for representations of quivers. *Compositio Math.* **130** (2002), 225–239.

[9] —, Normality of Marsden-Weinstein reductions for representations of quivers. *Math. Ann.* **325** (2003), 55–79.

[10] —, On matrices in prescribed conjugacy classes with no common invariant subspace and sum zero. *Duke Math. J.* **118** (2003), 339–352.

[11] —, Indecomposable parabolic bundles and the existence of matrices in prescribed conjugacy class closures with product equal to the identity. *Inst. Hautes Études Sci. Publ. Math.* **100** (2004), 171–207.

[12] —, Kac's Theorem for weighted projective lines. Preprint; math.AG/0512078.

[13] —, The Deligne-Simpson problem. In preparation.

[14] Crawley-Boevey, W., Etingof, P., Ginzburg, V., Noncommutative geometry and quiver algebras. Preprint; math.AG/0502301.

[15] Crawley-Boevey, W., Holland, M. P., Noncommutative deformations of Kleinian singularities. *Duke Math. J.* **92** (1998), 605–635.

[16] Crawley-Boevey, W., Shaw, P., Multiplicative preprojective algebras, middle convolution and the Deligne-Simpson problem. *Adv. Math.* **201** (2006), 180–206.

[17] Crawley-Boevey, W., Van den Bergh, M., Absolutely indecomposable representations and Kac-Moody Lie algebras (with an appendix by Hiraku Nakajima). *Invent. Math.* **155** (2004), 537–559.

[18] Etingof, P., Oblomkov, A., Rains, E., Generalized double affine Hecke algebras of rank 1 and quantized Del Pezzo surfaces. Preprint; math.QA/0406480.

[19] Gabriel, P., Unzerlegbare Darstellungen, I. *Manuscripta Math.* **6** (1972), 71–103.

[20] Geigle, W., Lenzing, H., A class of weighted projective curves arising in representation theory of finite dimensional algebras. In *Singularities, representations of algebras, and vector bundles* (Lambrecht, 1985), ed. by G.-M. Greuel and G. Trautmann, Lecture Notes in Math. 1273, Springer-Verlag, Berlin 1987, 265–297.

[21] Hall, P., The algebra of partitions. In *Proceedings of the Fourth Canadian Mathematical Congress* (Banff, 1957), University of Toronto Press, Toronto 1959, 147–159.

[22] Happel, D., *Triangulated categories in the representation theory of finite-dimensional algebras*. London Math. Soc. Lecture Note Ser. 119, Cambridge University Press, Cambridge 1988.

[23] Happel, D., Ringel, C. M., Tilted algebras. *Trans. Amer. Math. Soc.* **274** (1982), 399–443.

[24] Holland, M. P., Quantization of the Marsden-Weinstein reduction for extended Dynkin quivers. *Ann. Sci. École Norm. Sup.* (4) **32** (1999), 813–834.

[25] Hubery, A., From triangulated categories to Lie algebras: A theorem of Peng and Xiao. *Proceedings of the Workshop on Representation Theory of Algebras and related Topics* (Querétaro, 2004), ed. by J. De la Peña and R. Bautista, to appear.

[26] Kac, V. G., Infinite root systems, representations of graphs and invariant theory. *Invent. Math.* **56** (1980), 57–92.

[27] —, Root systems, representations of quivers and invariant theory. In *Invariant theory* (Montecatini, 1982), ed. by F. Gherardelli, Lecture Notes in Math. 996, Springer-Verlag, Berlin 1983, 74–108.

[28] Kostov, V. P., The Deligne-Simpson problem—a survey. *J. Algebra* **281** (2004), 83–108.

[29] Kronheimer, P. B., The construction of ALE spaces as hyper-Kähler quotients. *J. Differential Geom.* **29** (1989), 665–683.

[30] Le Bruyn, L., Noncommutative smoothness and coadjoint orbits. *J. Algebra* **258** (2002), 60–70.

[31] Lenzing, H., Representations of finite dimensional algebras and singularity theory. In *Trends in ring theory* (Miskolc, Hungary, 1996), ed. by V. Dlab and L. Márki, CMS Conf. Proc. 22, Amer. Math. Soc., Providence, RI, 1998, 71–97.

[32] Lenzing, H., Meltzer, H., Sheaves on a weighted projective line of genus one, and representations of a tubular algebra. In *Representations of algebras* (Ottawa, 1992), ed. by V. Dlab and H. Lenzing, CMS Conf. Proc. 14, Amer. Math. Soc., Providence, RI, 1993, 313–337.

[33] Lin, Y., Peng, L., Elliptic Lie algebras and tubular algebras. *Adv. Math.* **196** (2005), 487–530.

[34] Lusztig, G., Intersection cohomology methods in representation theory. In *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. I, The Mathematical Society of Japan, Tokyo, Springer-Verlag, Tokyo, 1991, 155–174.

[35] Moody, R. V., Rao, S. Eswara, Yokonuma, T., Toroidal Lie algebras and vertex representations. *Geom. Dedicata* **35** (1990), 283–307.

[36] Nakajima, H., Geometric construction of representations of affine algebras. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. I, Higher Ed. Press, Beijing 2002, 423–438.

[37] Neto, O., Silva, F. C., Singular regular differential equations and eigenvalues of products of matrices. *Linear Multilin. Algebra* **46** (1999), 145–164.

[38] Peng, L., Xiao, J., Triangulated categories and Kac-Moody algebras. *Invent. Math.* **140** (2000), 563–603.

[39] Reiten, I., Tilting theory and quasitilted algebras. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 109–120.

[40] Riedtmann, C., Algebren, Darstellungsköcher, Überlagerungen und zurück. *Comment. Math. Helv.* **55** (1980), 199–224.

[41] Ringel, C. M., Indecomposable representations of finite-dimensional algebras. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 425–436.

[42] —, *Tame algebras and integral quadratic forms*. Lecture Notes in Math. 1099, Springer-Verlag, Berlin 1984.

[43] —, Hall algebras. In *Topics in algebra* (Warsaw, 1988), ed. by S. Balcerzyk et al., Banach Center Publ. 26, Part 1, PWN, Warsaw 1990, 433–447.

[44] —, Hall algebras and quantum groups. *Invent. Math.* **101** (1990), 583–591.

[45] —, The preprojective algebra of a quiver. In *Algebras and modules, II* (Geiranger, 1996), ed. by I. Reiten et al., CMS Conf. Proc. 24, Amer. Math. Soc., Providence, RI, 1998, 467–480.

[46] Schiffmann, O., Noncommutative projective curves and quantum loop algebras. *Duke Math. J.* **121** (2004), 113–168.

[47] Steinitz, E., Zur Theorie der Abel'schen Gruppen. *Jahresber. DMV* **9** (1901), 80–85.

[48] Van den Bergh, M., Double Poisson algebras. Preprint; math.QA/0410528.

[49] Weil, A., Generalization de fonctions abeliennes. *J. Math. Pures Appl.* **17** (1938), 47–87.

Department of Pure Mathematics, University of Leeds, Leeds LS2 9JT, UK
E-mail: w.crawley-boevey@leeds.ac.uk

Zeta functions of groups and rings

Marcus du Sautoy and Fritz Grunewald

Abstract. We report on progress and problems concerning the analytical behaviour of the zeta functions of groups and rings. We also describe how these generating functions are special cases of adelic cone integrals for which our results hold.

Mathematics Subject Classification (2000). Primary 20E07, 11M41; Secondary 20F18, 17B30, 11M45.

Keywords. Zeta functions, nilpotent groups, rings, elliptic curves, analytic continuation, functional equations.

1. Introduction

Whenever a counting problem pertaining to some mathematical object Λ produces a sequence of non-negative integers $a_\Lambda(n)$ $(n = 1, 2, \dots)$ we can hope to gain information by incorporating our sequence into a generating function. There are various ways of doing this, for example as coefficients of power series, sums representing automorphic functions and Dirichlet series. Sometimes there is a natural choice of a generating function dictated by the recursive properties of the sequence $a_\Lambda(n)$. We report here on counting problems where the choice of a Dirichlet series seems to be appropriate.

We consider first two counting problems relating to a finitely generated group G. Write $[G : H]$ for the index of a subgroup $H \leq G$ and let

$$a_G^<(n) := |\{\, H \leq G \mid [G : H] = n \,\}|, \quad a_G^\lhd(n) := |\{\, H \lhd G \mid [G : H] = n \,\}| \quad (1)$$

be the number of subgroups or normal subgroups of index precisely n in G. The numbers $a_G^<(n)$ all being finite we call

$$\zeta_G^<(s) := \sum_{n=1}^{\infty} a_G^<(n)\, n^{-s} = \sum_{H \leq_f G} [G : H]^{-s} \quad (2)$$

the *subgroup zeta function* of G. The symbol $H \leq_f G$ indicates that the summation is over all subgroups H of finite index in G. Similarly, we define

$$\zeta_G^\lhd(s) := \sum_{n=1}^{\infty} a_G^\lhd(n)\, n^{-s} = \sum_{H \lhd_f G} [G : H]^{-s} \quad (3)$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

to be the *normal subgroup zeta function* of G. When we intend to address both types of zeta functions simultaneously we write $\zeta_G^*(s)$ for $\zeta_G^<(s)$ or $\zeta_G^\lhd(s)$.

For the second type of counting problem we consider a ring R, which is for our purposes an abelian group R carrying a biadditive product. Let us write $S \leq R$ if S is a subring of R and $\mathfrak{a} \trianglelefteq R$ if $\mathfrak{a}$ is a left ideal of R. Let

$$a_R^<(n) := |\{\, S \leq R \mid [R : S] = n \,\}|, \quad a_R^\lhd(n) := |\{\, \mathfrak{a} \trianglelefteq R \mid [R : \mathfrak{a}] = n \,\}| \tag{4}$$

be the numbers of these subobjects of R which have index $n \in \mathbb{N}$ in the additive group of R. The numbers counting subrings are finite if the additive group of R is finitely generated. The numbers counting ideals are all finite under the weaker hypothesis that the ring R is finitely generated. Given these circumstances define the *subring zeta function* or the *ideal zeta function* to be respectively

$$\zeta_R^<(s) := \sum_{n=1}^{\infty} a_R^<(n)\, n^{-s}, \quad \zeta_R^\lhd(s) := \sum_{n=1}^{\infty} a_R^\lhd(n)\, n^{-s}. \tag{5}$$

Again we write $\zeta_R^*(s)$ for $\zeta_R^<(s)$ or $\zeta_R^\lhd(s)$.

While the study of the zeta functions of a finitely generated group was only begun in [27], the ideal zeta function of a ring has the Riemann zeta function ($R = \mathbb{Z}$) or more generally the Dedekind zeta function of the ring of integers in a number field as special cases (see [32]).

We wish to consider the Dirichlet series (2), (3), (5) not only as formal sums but as series converging in a non-empty subset of the complex numbers. By general theory this subset may be taken to be a right half-plane. In fact, this convergence condition will be satisfied if and only if the coefficients in the series (2), (3), (5) grow at most polynomially in n, more precisely if and only if there are $t,\ c^* \in \mathbb{R}$ such that $a_G^*(n) \leq c^* n^t$ respectively $a_R^*(n) \leq c^* n^t$ holds for all $n \in \mathbb{N}$. In this case we will say that G has *polynomial subgroup* or *normal subgroup growth*, the ring R will be said to have *polynomial subring* or *ideal growth*. For finitely generated groups G there is the following beautiful characterisation of this property by A. Lubotzky, A. Mann and D. Segal (see [36]).

Theorem 1.1. *Let G be a finitely generated residually finite group. Then G has polynomial subgroup growth if and only if G has a subgroup of finite index which is soluble and of finite rank.*

A group G is called residually finite if for every non-trivial $g \in G$ there is a subgroup H of finite index in G with $g \notin H$. Of course, this assumption is natural for Theorem 1.1 to hold. A group G is said to be of finite rank $r \in \mathbb{N}$ if every finitely generated subgroup of G can be generated by at most r elements.

In the following we shall assume that

- either $\Lambda = G$ is an infinite finitely generated torsion-free nilpotent group,
- or $\Lambda = R$ is a ring with additive group isomorphic to $\mathbb{Z}^d$ for some $d \in \mathbb{N}$.

We shall denote the set of isomorphism classes of such objects by $\mathcal{T}$.

Finitely generated torsion-free nilpotent groups satisfy the growth condition in Theorem 1.1. Their classification up to isomorphism is intimately connected with the reduction theory for arithmetic groups (see [24]). In addition there are connections between special classes of nilpotent groups and certain diophantine problems including the question of equivalence classes of integral quadratic forms ([25]). See [26] for a panorama of finitely generated torsion-free nilpotent groups. The class of rings in $\mathcal{T}$ contains all rings of integers in number fields and also (for example) the integer versions of the simple Lie algebras over $\mathbb{C}$.

For $\Lambda \in \mathcal{T}$ the zeta functions share a number of features in common with the Dedekind zeta function of a number field. Before we report the story let us mention some examples. Considering $\mathbb{Z}^d$ ($d \in \mathbb{N}$) as a direct product of infinite cyclic groups we find $\zeta^*_{\mathbb{Z}^d}(s) = \zeta(s)\zeta(s-1)\dots\zeta(s-d+1)$ where

$$\zeta(s) = \zeta^<_{\mathbb{Z}}(s) = \zeta^\lhd_{\mathbb{Z}}(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p \frac{1}{1-p^{-s}}$$

is the Riemann zeta function. A more elaborate example concerns the discrete Heisenberg group H_3, that is the group of strictly upper triangular 3×3-matrices with integer entries. The group H_3 is a torsion-free, nilpotent group of class 2 generated by two elements. The following formulas are proved in [46], see also [27].

$$\zeta^<_{H_3}(s) = \frac{\zeta(s)\zeta(s-1)\zeta(2s-2)\zeta(2s-3)}{\zeta(3s-3)}, \quad \zeta^\lhd_{H_3}(s) = \zeta(s)\zeta(s-1)\zeta(3s-2). \tag{6}$$

For an interesting example of the zeta function of a ring we can consider $sl_2(\mathbb{Z})$, the additive group of integer 2×2-matrices of trace 0 with the usual Lie bracket. The following formula was finally proved in [21] after contributions in [29], [5], [6]

$$\zeta^<_{sl_2(\mathbb{Z})}(s) = P(2^{-s})\frac{\zeta(s)\zeta(s-1)\zeta(2s-2)\zeta(2s-1)}{\zeta(3s-1)} \tag{7}$$

where $P(x)$ is the rational function $P = (1+6x^2-8x^3)/(1-x^3)$.

All these examples of zeta functions of members of $\mathcal{T}$ have three distinctive properties (evident from the formulas given):

- they converge in some right halfplane of $\mathbb{C}$,
- they decompose similarly to the Riemann or Dedekind zeta function as an Euler product of some rational expression in p^{-s} taken over all primes p,
- they have a meromorphic continuation to $\mathbb{C}$.

We believe that these three properties already justify the name zeta function for the corresponding generating function. A fourth property of the Dedekind zeta function, the global functional equation (see [32]) is hardly conceivable looking at formulas (6), (7).

Let us define now what will be the Euler factors of the zeta function of a general $\Lambda \in \mathcal{T}$. For a prime p we set:

$$\zeta^*_{\Lambda,p}(s) := \sum_{k=0}^{\infty} a^*_\Lambda(p^k)\, p^{-ks}. \tag{8}$$

This expression can be considered as a function in the variable $s \in \mathbb{C}$ or equally as a power series in p^{-s}.

In [27] the following theorem is established.

Theorem 1.2. *For $\Lambda \in \mathcal{T}$ the following hold.*

(i) *The Dirichlet series $\zeta^*_\Lambda(s)$ converges in some right half-plane of $\mathbb{C}$.*

(ii) *The Dirichlet series $\zeta^*_\Lambda(s)$ decomposes as an Euler product*

$$\zeta^*_\Lambda(s) = \prod_p \zeta^*_{\Lambda,p}(s), \tag{9}$$

where the product is to be taken over all primes p.

(iii) *The power series $\zeta^*_{\Lambda,p}(s)$ are rational functions in p^{-s}. That is, for each prime p there are polynomials $Z^*_p,\ N^*_p \in \mathbb{Z}[x]$ such that $\zeta^*_{\Lambda,p}(s) = Z^*_p(p^{-s})/N^*_p(p^{-s})$ holds. The polynomials $Z^*_p,\ N^*_p$ can be chosen to have bounded degree as p varies.*

An explicit determination of the local Euler factors (that is of the polynomials $Z^*_p,\ N^*_p$) of the zeta functions has been carried out in many cases including infinite families of examples. The methods range from ingenious elementary arguments to the use of algebraic geometry (resolving singularities). In several cases computer assistance was used (see [51]). See Section 6 for a selection of these examples. The database [52] collects comprehensive information on many examples treated so far. In very few cases the zeta function could be described by a closed formula in terms of the Riemann zeta function like in (6) or (7).

Note that, as a consequence of Theorem 1.2, the series (2), (3) and (5) converge to holomorphic functions in some right half-plane of $\mathbb{C}$. In fact the coefficients in (2), (3) and (5) are non-negative, hence by a well known theorem of E. Landau there is $\alpha \in \mathbb{R} \cup \{-\infty\}$ such that the series in question converges (absolutely and locally uniformly) for $s \in \mathbb{C}$ with $\mathrm{Re}(s) > \alpha$ and diverges if $\mathrm{Re}(s) < \alpha$. This α is called the *abscissa of convergence* of the series.

We wish to report the following theorem which collects together the main results of [13].

Theorem 1.3. *For $\Lambda \in \mathcal{T}$ the following hold.*

(i) *The abscissa of convergence α^*_Λ of $\zeta^*_\Lambda(s)$ is a rational number.*

(ii) *There is a $\delta > 0$ such that $\zeta^*_\Lambda(s)$ can be meromorphically continued to the region $\{\, s \in \mathbb{C} \mid \mathrm{Re}(s) > \alpha^*_\Lambda - \delta \,\}$.*

(iii) *The line $\{\, s \in \mathbb{C} \mid \mathrm{Re}(s) = \alpha^*_\Lambda \,\}$ contains at most one pole of $\zeta^*_\Lambda(s)$ (at the point $s = \alpha^*_\Lambda$).*

We define b^*_Λ to be the order of the pole of $\zeta^*_\Lambda(s)$ in $s = \alpha^*_\Lambda$. Using another theorem of E. Landau we find $b^*_\Lambda \geq 1$. Theorem 1.3 has as an immediate consequence:

Corollary 1.1. *Let Λ be in $\mathcal{T}$. Let*

$$s^*_\Lambda(N) := \sum_{n=1}^{N} a^*_\Lambda(n) \quad (N \in \mathbb{N}) \tag{10}$$

be the summatory function of the counting function a^*_Λ. We have

$$s^*_\Lambda(N) \sim c^*_\Lambda N^{\alpha^*_\Lambda} \log(N)^{b^*_\Lambda - 1} \tag{11}$$

with $c^*_\Lambda \in \mathbb{R}$.

The formula (11) means that the right hand side divided by the left hand side tends to 1 as N tends to infinity. The corollary follows from Theorem 1.3 using Tauber theory (see [38]). Note that the third property of the zeta function is essential for this application. Note also that c^*_Λ is equal to the lowest coefficient of the Laurent series representing ζ^*_Λ near the pole in $s = \alpha^*_\Lambda$.

Having defined the new invariants $\alpha^*_\Lambda \in \mathbb{Q}$, $b^*_\Lambda \in \mathbb{N}$ and $c^*_\Lambda \in \mathbb{R}$ for every Λ in $\mathcal{T}$ we are lead to

Problem 1.1. Relate $\alpha^*_\Lambda, b^*_\Lambda, c^*_\Lambda \in \mathbb{R}$ to structural properties of Λ.

This problem is solved when Λ is the ring of integers of a number field for the ideal zeta function. In this case $\alpha^\triangleleft_\Lambda = b^\triangleleft_\Lambda = 1$ and the value of $c^\triangleleft_\Lambda$ is given by Dirichlet's class number formula (see [32]). In the general case we have only the very scarce information reported in later sections. The following asymptotic relations can be read off from formulas (6) and (7), they reveal the values of our invariants.

$$s^<_{H_3}(N) \sim \frac{\zeta(2)^2}{2\zeta(3)} N^2 \log(N), \quad s^<_{sl_2(\mathbb{Z})}(N) \sim \frac{20\zeta(2)^2\zeta(3)}{31\zeta(5)} N^2 \tag{12}$$

The examples above illustrate that α^*_Λ can often be any natural number. However examples described in Section 6 show that $5/2$ and $7/2$ are also possible values of α^*_Λ. Considering $\mathbb{Z}^d$ ($d \in \mathbb{N}$) as a ring we have $\zeta^\triangleleft_{\mathbb{Z}^d}(s) = \zeta(s)^d$ and hence $b^\triangleleft_{\mathbb{Z}^d} = d$. Examples from Section 6 illustrate that b^*_Λ can take the values 1, 2, 3, 4. In fact, the Heisenberg group H_3 has $b^<_{H_3} = 2$. This is most of the knowledge we so far have on

Problem 1.2. What is the range of the pairs $(\alpha^*_\Lambda, b^*_\Lambda)$ as Λ varies over $\mathcal{T}$?

Problem 1.2 has many more concrete variants, let us mention one of them. Define $\mathbf{S}^*_{\text{group}} := \{\alpha^*_G\} \subset \mathbb{R}$ to be the set of abscissas of convergence of the subgroup or normal subgroup zeta functions as G varies over all finitely generated torsion-free nilpotent groups. Define $\mathbf{S}^*_{\text{ring}} := \{\alpha^*_R\} \subset \mathbb{R}$ similarly as R varies over all rings in $\mathcal{T}$. Let us briefly explain the proof of

Proposition 1.1. *The set* $\mathbf{S}^{<}_{\text{group}} \subset \mathbb{R}$ *is discrete, that is below any real number there are only finitely many members of* $\mathbf{S}^{<}_{\text{group}}$.

Let G be a torsion-free nilpotent group and G^{ab} its abelianisation. Let $h(G)$ be the Hirsch-length of G, that is the maximal number of infinite cyclic factors appearing in a composition series of G. A simple argument shows $h(G^{\text{ab}}) \leq \alpha^*_G \leq h(G)$. These two inequalities have been improved in several directions (see [27] and [42]). An unpublished result of D. Segal gives the lower bound

$$\big(3 - \sqrt{2}\big)h(G) - \frac{1}{2} \leq \alpha^{<}_G. \tag{13}$$

The main result of [4] implies that once we fix the Hirsch-length of G there is a universal denominator which all denominators of local zeta functions of nilpotent groups of that Hirsch-length have to divide. The results of Sections 2 and 3 together with (13) prove Proposition 1.1. The results of [4] apply also to the normal subgroup zeta function and to the zeta functions of rings, but a replacement for (13) has not been found. So we raise

Problem 1.3. Are the sets $\mathbf{S}^{\lhd}_{\text{group}}$, $\mathbf{S}^{<}_{\text{ring}}$ and $\mathbf{S}^{\lhd}_{\text{ring}}$ discrete? If not, what are their accumulation points?

Are $\Lambda_1,\ \Lambda_2 \in \mathcal{T}$ isomorphic if their zeta functions agree? Questions of this nature are traditionally called isospectrality problems. Examples of non-isomorphic rings of integers $R_1,\ R_2$ in number fields with $\zeta^{\lhd}_{R_1}(s) = \zeta^{\lhd}_{R_2}(s)$ are contained in [44]. The two finitely generated nilpotent groups (of class 2) G_1, G_2 described in Example 4 of Section 6 are not isomorphic but have isomorphic profinite completions. Hence $\zeta^{<}_{G_1}(s) = \zeta^{<}_{G_2}(s)$, $\zeta^{\lhd}_{G_1}(s) = \zeta^{\lhd}_{G_2}(s)$ both hold. These examples show that the isospectrality problem in general has a negative answer. But there remains:

Problem 1.4. Suppose that $\zeta^*_{\Lambda_1}(s) = \zeta^*_{\Lambda_2}(s)$ holds for both or at least one of the possibilities $* \in \{<, \lhd\}$ for $\Lambda_1,\ \Lambda_2 \in \mathcal{T}$. Which structural invariants of Λ_1 and Λ_2 are the same? For example, are the profinite completions of Λ_1 and Λ_2 isomorphic?

For a more extensive discussion of isospectrality problems see [12]. This paper also contains an example of a group G which satisfies $\zeta^{<}_G(s) = \zeta^{<}_{\mathbb{Z}^2}$ but which does not have the same profinite completion as $\mathbb{Z}^2$. The group G is one of the plane crystallographic groups, it has $\mathbb{Z}^2$ as a subgroup of index 2 but it is not nilpotent.

In this survey we mainly discuss properties of the zeta functions of groups and rings. There are many topics not treated here, see [20] and [10] for relations to other subjects. Connections to the by now vast field of subgroup growth are not treated here. For this see the surveys [34] and [35].

In Sections 2, 3 we describe the proofs of Theorems 1.2 and 1.3. As a first step the Euler factors $\zeta^*_{\Lambda,p}(s)$ are described as certain p-adic integrals. These are evaluated by the methods of p-adic integration. Having obtained explicit formulas we multiply the (global) Euler product by an Artin L-function to enlarge its region of

convergence. Section 4 discusses the variation with p of the Euler factors. Certain functional equations of the Euler factors are the subject of Section 5. Section 6 contains examples. In Section 7 we describe variations on the zeta function theme.

2. p-adic formalism

While the proof of the first two items of Theorem 1.2 is elementary, the third requires an expression for the local Euler factor $\zeta^*_{\Lambda,p}(s)$ (p a prime) of the zeta function $\zeta^*_\Lambda(s)$ in terms of a certain p-adic integral. We shall briefly explain this procedure in the case when $\Lambda = R \in \mathcal{T}$ is a ring and $* = \lhd$. For more details see [27] Section 3 or [13].

Let p be a prime. We write $\mathbb{Q}_p$ for the field of p-adic numbers and $\mathbb{Z}_p$ for its ring of integers. For $x \in \mathbb{Q}_p$ we define $v_p(x)$ to be the p-adic valuation of x and $|x|_p$ to be the normalised p-adic absolute value. We write $\mathrm{Tr}_d(\mathbb{Z})$, $\mathrm{Tr}_d(\mathbb{Z}_p)$ ($d \in \mathbb{N}$) for the space of upper triangular $d \times d$-matrices with entries in $\mathbb{Z}$ respectively $\mathbb{Z}_p$. We think of $\mathrm{Tr}_d(\mathbb{Z}_p)$ being identified with $\mathbb{Z}_p^{d(d+1)/2}$.

Let $R \in \mathcal{T}$ be a ring (with additive group isomorphic to $\mathbb{Z}^d$) and p a prime. We fix a $\mathbb{Z}$-basis of R. Analysing the conditions for the rows of an upper triangular matrix (in $\mathrm{Tr}_d(\mathbb{Z})$) to be a triangular basis of an ideal in R, we find polynomials

$$f_1,\ g_1, \ldots,\ f_l, \quad g_l \in \mathbb{Z}[x_{11}, \ldots, x_{dd}] \tag{14}$$

such that

$$\mathcal{M}^\lhd(R) := \{\, x \in \mathrm{Tr}_d(\mathbb{Z}) \mid f_1(x) \mid g_1(x), \ldots,\ f_l(x) \mid g_l(x) \,\} \tag{15}$$

is exactly the set of upper triangular matrices with entries in $\mathbb{Z}$ for which the rows generate an ideal in R. Here we write $a \mid b$ if the integer a divides the integer b. We now use our $\mathbb{Z}$-basis of R also as a $\mathbb{Z}_p$-basis of $\mathbb{Z}_p \otimes_\mathbb{Z} R$. We conclude that

$$\mathcal{M}^\lhd(R,\, p) := \{\, x \in \mathrm{Tr}_d(\mathbb{Z}_p) \mid v_p(f_i(x)) \le v_p(g_i(x)) \text{ for } i = 1, \ldots, l \,\} \tag{16}$$

is exactly the set of upper triangular matrices with entries in $\mathbb{Z}_p$ for which the rows additively generate an ideal in $\mathbb{Z}_p \otimes_\mathbb{Z} R$.

The map $\mathfrak{a} \to \mathfrak{a} \cap R$ sets up a one to one correspondence between the ideals of index p^n ($n \in \mathbb{N}$) in $\mathbb{Z}_p \otimes_\mathbb{Z} R$ and ideals of the same index in R. An exercise in p-adic integration shows that

$$\zeta^\lhd_{R,p}(s) = (1 - p^{-1})^{-d} \int_{\mathcal{M}^\lhd(R,p)} |x_{11}|_p^{s-n} |x_{22}|_p^{s-n+1} \ldots |x_{dd}|_p^{s-1}\, dx \tag{17}$$

holds for every prime p with dx the normalised Haar measure on $\mathrm{Tr}_d(\mathbb{Z}_p) = \mathbb{Z}_p^{d(d+1)/2}$. The same approach applies to the subring zeta function of a ring $R \in \mathcal{T}$. See Section 3 of [27] or Section 5 of [13] for more details.

A similar, slightly more elaborate, analysis yields polynomials (14) in $d(d+1)/2$ variables such that formula (17) holds in case $\Lambda = G$ is a finitely generated torsion-free nilpotent group. For more details see Section 2 of [27] and Section 5 of [13]. Here, due to the use of Lie ring methods, finitely many primes have to be excluded. The natural number d has to be taken to be the Hirsch-length of G.

The proof in [27] of the rationality of these p-adic integrals relied on observing that $\mathcal{M}^*(\Lambda, p)$ are definable subsets in the language of fields. One can then apply a theorem of Denef [1] which establishes the rationality of definable p-adic integrals. Denef's proof relies on an application of Macintyre's quantifier elimination for the theory of $\mathbb{Q}_p$ which simplifies in a generally mysterious way the description of definable subsets like $\mathcal{M}^*(\Lambda, p)$. In the next section we shall report on a concrete formula computing integrals like (17) which replaces the use of the model theoretic black box in the proof of the rationality.

3. p-adic and adelic cone integrals

We define here certain Euler products with factors given by p-adic integrals which are generalized versions of the p-adic integrals occurring in formula (17). We then analyze the analytical properties of these Euler products.

Let m be a natural number. A collection of polynomials

$$\mathcal{D} = (f_0, g_0; f_1, g_1, \ldots, f_l, g_l) \quad (f_0, g_0, f_1, g_1, \ldots, f_l, g_l \in \mathbb{Q}[x_1, \ldots, x_m]) \tag{18}$$

is called cone integral data. We associate to $\mathcal{D}$ the following closed subset of $\mathbb{Z}_p^m$ (p a prime)

$$\mathcal{M}(\mathcal{D}, p) := \{ x \in \mathbb{Z}_p^m \mid v_p(f_i(x)) \leq v_p(g_i(x)) \text{ for } i = 1, \ldots, l \} \tag{19}$$

and a p-adic integral with conventions as in Section 2:

$$Z_{\mathcal{D}}(s, p) = \int_{\mathcal{M}(\mathcal{D}, p)} |f_0(x)|_p^s \, |g_0(x)|_p \, dx. \tag{20}$$

Note that Section 2 shows that the local zeta functions of the $\Lambda \in \mathcal{T}$ are special cases of the $Z_{\mathcal{D}}(s, p)$. The p-adic integral (20) is easily seen to exist for $s \in \mathbb{C}$ with sufficiently large real part. It can be expressed as a power series $Z_{\mathcal{D}}(s) = \sum_{i=0}^{\infty} a_{p,i} \, p^{-is}$ with non-negative integer coefficients and $a_{p,0} \neq 0$. In fact, a result of Denef [1] says that the power series in (20) is rational in p^{-s}. Given the cone integral data $\mathcal{D}$ we can define $Z_{\mathcal{D}}(s, p)$ for every prime p. We use this to define an Euler product

$$Z_{\mathcal{D}}(s) = \prod_p (a_{p,0}^{-1} \cdot Z_{\mathcal{D}}(s, p)) \tag{21}$$

which we call the *global* or *adelic cone integral*. In fact, with appropriate normalisation of measures, $Z_{\mathcal{D}}(s)$ can be defined as an adelic integral (see [39] for a special case).

Special cases of the p-adic integrals (20) appear in [28]. We use an adaptation of a method to calculate p-adic integrals from [1] and [2] to show:

Proposition 3.1. *Let $\mathcal{D} = (f_0, g_0; f_1, g_1, \ldots, f_l, g_l)$ be cone integral data. Define the polynomial $F(x) := \prod_{i=0}^{l} f_i(x)g_i(x)$. Let (Y, h) be a resolution of singularities over $\mathbb{Q}$ of F. Let E_i $(i \in T)$ be the irreducible components defined over $\mathbb{Q}$ of the reduced scheme $(h^{-1}(D))_{\mathrm{red}}$ where $D = \mathrm{Spec}(\mathbb{Q}[x]/F)$. Then the following hold.*

(i) *There exist rational functions $P_I(X, Y) \in \mathbb{Q}(X, Y)$ for each $I \subset T$ with the property that for almost all p:*

$$Z_{\mathcal{D}}(s) = \sum_{I \subset T} c_{p,I}\, P_I(p, p^{-s}) \tag{22}$$

where $c_{p,I} = |\{a \in \bar{Y}(\mathbb{F}_p) \mid a \in E_i \text{ if and only if } i \in I\}|$ and $\bar{Y}$ is the reduction mod p of the scheme Y.

(ii) *There is a closed polyhedral cone $\mathfrak{C} \subset \mathbb{R}^t_{\geq 0}$ where $t = |T|$ and a decomposition of $\mathfrak{C}$ into open simplicial pieces which we denote by R_k $(k \in \{0, 1, \ldots, w\})$. We arrange that $R_0 = (0, \ldots, 0)$ and $R_1, \ldots, R_q$ are the one-dimensional pieces. For each $k \in \{0, 1, \ldots, w\}$ let $M_k \subset \{1, \ldots, q\}$ denote the one-dimensional pieces in the closure of R_k. Then there are positive integers A_j, B_j for $j \in \{1, \ldots, q\}$ such that for almost all primes p:*

$$Z_{\mathcal{D}}(s) = \sum_{k=0}^{w} (p-1)^{I_k}\, p^{-m}\, c_{p,I_k} \prod_{j \in M_k} \frac{p^{-(A_j s + B_j)}}{1 - p^{-(A_j s + B_j)}} \tag{23}$$

where I_k is the subset of T defined so that $i \in T \setminus I_k$ if and only if the i-th coordinate is zero for all elements of R_k.

The study of p-adic integrals like (20) has been initiated by J. Igusa. His fundamental results are documented in [28]. The references in [28] provide access to the vast literature on this subject. Previous to the results documented here the global or adelic versions (21) have only received attention in special cases (see [39]). Using various methods from analytic number theory and arithmetic geometry we show in [13] that Proposition 3.1 implies:

Corollary 3.1. *Let $\mathcal{D} = (f_0, g_0; f_1, g_1, \ldots, f_l, g_l)$ be cone integral data. Suppose $Z_{\mathcal{D}}(s)$ is not the constant function.*

(i) *The abscissa of convergence $\alpha = \alpha_{\mathcal{D}}$ of $Z_{\mathcal{D}}(s) = \sum_{n=1}^{\infty} a_n n^{-s}$ is a rational number.*

(ii) *$Z_{\mathcal{D}}(s)$ has a meromorphic continuation to $\mathrm{Re}(s) > \alpha - \delta$ for some $\delta > 0$.*

(iii) *The line $\{ s \in \mathbb{C} \mid \mathrm{Re}(s) = \alpha_\Lambda \}$ contains only one pole of $\zeta_{\mathcal{D}}(s)$ at $s = \alpha_\Lambda$.*

In fact, we multiply $Z_{\mathcal{D}}(s)$ by the Artin L-function corresponding to the permutation representation of the absolute Galois group of $\mathbb{Q}$ on the irreducible components of the E_i $(i \in T)$ appearing in Proposition 3.1. Using the estimates of Hasse and

Weil for the number of points on algebraic varieties over finite fields on the c_{p,I_k} in formula (23) we can analyse the analytic properties of the product of $Z_{\mathcal{D}}(s)$ with its Artin L-function near the abscissa of convergence of $Z_{\mathcal{D}}(s)$.

Theorem 1.3 is a consequence of this corollary together with the discussion in Section 2.

Following classical analytic number theory it is natural to ask how far the adelic cone integrals $Z_{\mathcal{D}}(s)$ can be meromorphically continued to the left. The analysis of special cases shows that often natural boundaries arise as one continues to the left (see [11]). That is, in these cases, poles or zeroes of the continued function accumulate densely to the points of a vertical line in $\mathbb{C}$. Beyond this line no continuation is possible. The following problems seem to be of interest.

Problem 3.1. Find all cone integral data $\mathcal{D}$ such that $\zeta_{\mathcal{D}}(s)$ has a meromorphic continuation to $\mathbb{C}$, or at least give sufficient conditions for this to happen.

Problem 3.2. Find all $\Lambda \in \mathcal{T}$ such that $\zeta_{\Lambda}(s)$ has a meromorphic continuation to $\mathbb{C}$, or at least give sufficient conditions for this to happen.

Problem 3.3. Show that either $\zeta_{\mathcal{D}}(s)$ has a meromorphic continuation to $\mathbb{C}$ or that there is some rational number $\beta_{\mathcal{D}}$ such that the line $\{ s \in \mathbb{C} \mid \mathrm{Re}(s) = \beta_{\Lambda} \}$ is a natural boundary.

In [14] and [16] we attempt, partially successfully, to replace the zeta function by a ghost zeta function which has more amenable analytic properties but which has Euler factors which are in a specific sense near to those of the original zeta function.

The process of continuation to the left ties up the Dirichlet series $\zeta_{\mathcal{D}}(s)$ with the zeta functions defined by A. Weil and R. Langlands for smooth $\mathbb{Q}$-defined projective algebraic varieties. Let us report on a special example. Let $y^2-x^3-ax-b\,(a,\ b \in \mathbb{Q})$ be a polynomial representing an elliptic curve E. Define

$$Z_{E,p}(s) := \int_{\mathbb{Z}_p^2} |y^2 - x^3 - ax - b|_p^s \, dx, \quad Z_E(s) := \prod_p (\lambda_p^{-1} Z_{E,p}(s)) \qquad (24)$$

with appropriate normalisation factors λ_p. We have shown in [17] that the Dirichlet series $Z_E(s)$ converges for $\mathrm{Re}(s) > 0$. Moreover when attempting to continue $Z_E(s)$ to the left, the symmetric power L-functions attached to E arise. It is conjectured that these symmetric power L-functions can all be meromorphically continued to $\mathbb{C}$.[1] If this is true then $Z_E(s)$ can be meromorphically continued to the region $\mathrm{Re}(s) > -3/2$. Results of J. P. Serre concerning the Sato–Tate conjecture for E then imply that the line $\mathrm{Re}(s) = -3/2$ is a natural boundary beyond which no continuation is possible.

[1] Note added in proof: these conjectures have recently been proved.

4. The local factors: variation with p

The behaviour of the local factors as we vary p is one of the other major problems in the field. If we consider formula (6) we easily see that

$$\zeta_{H_3,p}(s) = \frac{W_1(p, p^{-s})}{W_2(p, p^{-s})} \tag{25}$$

where W_1, W_2 can be given without reference to the prime p as polynomials $W_1(X, Y) = (1 - Y)(1 - XY)(1 - X^2Y^2)(1 - X^3Y^2)$, $W_2(X, Y) = 1 - X^3Y^3$. Groups with this property are said to have *uniform subgroup* or *normal subgroup growth*. In [27] it is proved that a finitely generated free nilpotent group of class 2 has both uniform subgroup and normal subgroup growth. As revealed in [7] the following problem ties up intimately with classification problems of finite p-groups, in particular with Higman's PORC-conjecture.

Problem 4.1. Show that every finitely generated free nilpotent group has both uniform subgroup and normal subgroup growth.

We can also consider a similar variety of problems for rings. We define *uniform subring* or *ideal growth* as above in the group case and raise

Problem 4.2. Show that the following Lie rings have uniform subring and ideal growth.

- Free nilpotent Lie rings of finite $\mathbb{Z}$-rank,
- $sl_n(\mathbb{Z})$ ($n \in \mathbb{N}$) or any other integer version of a simple Lie algebra over $\mathbb{C}$.

Let us now consider the Heisenberg group H_3 with entries from the ring of integers of a quadratic number field. The behaviour of the local factors of its zeta functions depends on how p behaves in the number field [27]. That is formulas like (25) hold, but finitely many pairs of polynomials are needed to describe the variation of the local factor of the zeta function with p. Groups with this property are said to have *finitely uniform subgroup* or *normal subgroup growth*. There is a similar concept in the case of rings.

For a long time this was the only type of variation with p which was known. Our explicit formula however takes the subject away from the behaviour of primes in number fields to the problem of counting points modulo p on a variety, a question which is in general wild and far from the uniformity predicted by all previous examples seen in [27]. Two papers [8] and [9] by the first author contain the following example of a class two nilpotent group of Hirsch length 9 whose zeta function depends on counting points mod p on the elliptic curve $y^2 = x^3 - x$. Define

$$G = \left\langle \begin{matrix} x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6, \\ y_1,\ y_2,\ y_3 \end{matrix} \;\middle|\; \begin{matrix} [x_1, x_4] = y_3,\ [x_1, x_5] = y_1, [x_1, x_6] = y_2, \\ [x_2, x_4] = y_2, [x_2, x_6] = y_1,\ [x_3, x_4] = y_1, \\ [x_3, x_5] = y_3 \end{matrix} \right\rangle \tag{26}$$

with the convention that commutators not mentioned are equal to 1. By [9] there exist rational functions $P_1(X, Y)$, $P_2(X, Y) \in \mathbb{Q}(X, Y)$ such that for almost all primes p

$$\zeta^{\triangleleft}_{G,p}(s) = P_1(p, p^{-s}) + |E(\mathbb{F}_p)|P_2(p, p^{-s}), \tag{27}$$

where E is the elliptic curve $y^2 = x^3 - x$. In [9] this formula is used to show that G is not finitely uniform. To see where the elliptic curve is hidden in the above presentation, take the determinant of the 3×3 matrix (a_{ij}) with entries $a_{ij} = [x_i, x_{j+3}]$ and you will get the projective version of E.

Formula (23) shows that the variation type with p of the Euler factors $\zeta_{\mathcal{D},p}(s)$ ($\mathcal{D}$ cone condition data) is that of functions counting points on a $\mathbb{Q}$-defined algebraic variety modulo primes p. But might their be further restrictions once we consider the Euler factors $\zeta_{\Lambda,p}(s)$ for $\Lambda \in \mathcal{T}$?

Problem 4.3. Let V be a $\mathbb{Q}$-defined algebraic variety. Is there $\Lambda_V \in \mathcal{T}$ such that there are rational functions $P_1(X, Y)$, $P_2(X, Y) \in \mathbb{Q}(X, Y)$ such that for almost all primes p

$$\zeta_{\Lambda_V,p}(s) = P_1(p, p^{-s}) + |V(\mathbb{F}_p)|P_2(p, p^{-s}) \tag{28}$$

holds?

The consideration of zeta functions obtained by motivic integration (see [18]) sheds some light on this new dialogue between groups and rings and questions of arithmetic geometry.

5. Functional equations of the local factors

There is another remarkable feature of many of the rational functions representing the local zeta function of nilpotent groups: they satisfy a certain palindromic symmetry. Let us explain this in the case of the normal subgroup zeta function of $F_{2,3}$, the free nilpotent group of class two on three generators. The group $F_{2,3}$ is torsion-free and has Hirsch-length 6. From [27] we know that

$$\zeta^{\triangleleft}_{F_{2,3},p}(s) = \left.\frac{1 + X^3Y^3 + X^4Y^3 + X^6Y^5 + X^7Y^5 + X^{10}Y^8}{(1-Y)(1-XY)(1-X^2Y)(1-X^8Y^5)(1-X^6Y^9)}\right|_{X=p,\, Y=p^{-s}} \tag{29}$$

holds for every prime p. Let us replace p by p^{-1} (and p^{-s} by p^s) in this expression. Indicating this replacement by $p \to p^{-1}$, we find:

$$\zeta^{\triangleleft}_{F_{2,3},p}(s)|_{p\to p^{-1}} = p^{15-9s}\zeta^{\triangleleft}_{F_{2,3},p}(s). \tag{30}$$

This phenomenon was found in all examples of all finitely generated nilpotent groups of class 2 and Lie rings of nilpotency class 2 where explicit computations have been done. We pose here the

Problem 5.1. Let G be a nilpotent group of class 2 and Hirsch-length h. Assume that the quotient of G modulo its center (which is abelian) has torsion-free rank m. Show that

$$\zeta^{<}_{G,p}(s)|_{p\to p^{-1}} = (-1)^d\, p^{\frac{h(h-1)}{2}-hs}\, \zeta^{<}_{G,p}(s), \tag{31}$$

$$\zeta^{\lhd}_{G,p}(s)|_{p\to p^{-1}} = (-1)^d\, p^{\frac{h(h-1)}{2}-(h+m)s}\, \zeta^{\lhd}_{G,p}(s) \tag{32}$$

hold for almost all primes p.

In [48] C. Voll is able to answer Problem 5.1 affirmatively for the special case of local zeta functions counting normal subgroups of torsion-free class 2 nilpotent groups which have a centre of $\mathbb{Z}$-rank 2 by giving explicit formulas for the local zeta functions in this case. C. Voll [49] and P. Paajanen [40], [43] and [41] have also confirmed the functional equation for the normal subgroup zeta function in more general settings by analysing the geometry of the Pfaffian hypersurface associated to presentations of class 2 nilpotent groups. Note however that the functional equation for zeta functions of nilpotent groups is not a completely general phenomenon. The Lie ring $\mathfrak{L}_W$ introduced in Example 3 of the next section has nilpotency class 3. The Euler factors of the ideal counting zeta function do not satisfy a functional equation, although the Euler factors of the subring zeta function do have such a symmetry.

Problem 5.1 should be seen in connection with a result of Denef and Meuser [2] who prove that the rational expression (in p^{-s}) corresponding to the Igusa-type p-adic integral

$$Z_{\{g_0,\,1\},p}(s) := \int_{\mathbb{Z}_p^m} |g_0(x)|_p^s\, dx \tag{33}$$

satisfy a functional equation if $g_0 \in \mathbb{Z}_p[x_1,\ldots,x_m]$ is absolutely irreducible and defines a smooth projective hypersurface over the finite field $\mathbb{F}_p$. A key role in their proof is played by the functional equation satisfied by the algebraic geometric zeta function for this hypersurface proved by A. Weil.

In [48] C. Voll uses the functional equation for the local zeta functions of elliptic curves to prove that the zeta function (27) of the nilpotent group encoding an elliptic curve in its presentation has a functional equation of the type predicted by (31). The paper [31] of B. Klopsch and C. Voll treats interesting new counting problems related to orthogonal and unitary groups over finite fields which arose in the study of functional equations.

The only counterexamples to the functional equations for zeta functions of groups and rings relate to counting normal subgroups in groups or ideals in rings. We therefore raise the following:

Problem 5.2. Let Λ be in $\mathcal{T}$. Show that there are rational numbers a, b and c such that

$$\zeta^{<}_{\Lambda,p}(s)|_{p\to p^{-1}} = (-1)^c\, p^{as+b}\, \zeta^{<}_{\Lambda,p}(s) \tag{34}$$

for almost all primes p.

In his thesis [51] L. Woodward analyses a general setting in which certain cone integrals are conjectured to have functional equations which could generalise the result of Denef and Meuser. The cone integral data has to satisfy what Woodward calls a homogeneity condition, namely $\deg(g_i) = \deg(f_i) + 1$ for $i = 1, \ldots, l$. The cone integrals describing zeta functions counting all subgroups or subrings satisfy this homogeneity condition in contrast to the normal subgroup and ideal zeta functions. Using results of Stanley on functional equations of polyhedral cones Woodward can prove the functional equation in the special case that all the polynomials of the cone integral data are monomials.

6. Examples

This section contains a brief description of the information obtained so far on the zeta functions of several series of Lie rings. We also describe the pair of finitely generated nilpotent groups which solves the isospectrality problem negatively.

Example 1. Let $\mathfrak{F}(2, n)$ be the free nilpotent Lie ring of class two on $n \in \mathbb{N}$ $(n \geq 2)$ generators. This Lie ring has $\mathbb{Z}$-rank $h(n) = n + n(n+1)/2$. Note that the Lie ring of the Heisenberg group H_3 is isomorphic to $\mathfrak{F}(2, 2)$. An explicit formula for the Euler factors of the ideal zeta function is given by C. Voll in [48] (see [40] for special cases). From these

$$\alpha^{\triangleleft}_{\mathfrak{F}(2,n)} = \max \left\{ n, \frac{\left(\frac{n(n-1)}{2} - j\right)(n+j)+1}{h(n)-j} \;\middle|\; j = 1, \ldots, \frac{n(n-1)}{2} - 1 \right\} \quad (35)$$

can be deduced. This formula shows that for $n \geq 5$, the abscissa of convergence of the global ideal zeta function is greater than n and is usually not an integer. However, sometimes it may just happen to be an integer. The only n in the range $5 \leq n \leq 200$ for which this happens is $n = 26$. Furthermore the ideal growth of $\mathfrak{F}(2, n)$ is uniform and the Euler factors $\zeta_{\mathfrak{F}(2,n),p}$ satisfy the functional equation of Problem 5.1 (see [50]).

Example 2. Let n be a natural number with $n \geq 2$. Define $\mathfrak{G}(n)$ to be the Lie ring

$$\mathfrak{G}(n) := \langle\, z,\ x_1, \ldots, x_{n-1},\ y_1, \ldots, y_{n-1} \mid [z, x_i] = y_i \ (i = 1, \ldots, n-1) \,\rangle. \quad (36)$$

Our convention again is that all commutators between the generators not mentioned are equal to 0. Hence $\mathfrak{G}(n)$ has nilpotency class two and $\mathbb{Z}$-rank $2n - 1$. D. Grenham [22] has determined explicit formulas for the ideal zeta functions of $\mathfrak{G}(n)$ for $n = 2, 3, 4, 5$. Let us report his formula in case $n = 4$. Define $W_4(X, Y)$ to be the rational function

$$W_4(X, Y) := \frac{1 + X^4Y^3 + X^5Y^3 + X^8Y^5 + X^9Y^5 + X^{18}Y^8}{(1-Y)(1-XY)(1-X^2Y)(1-X^3Y)(1-X^6Y^3)(1-X^{10}Y^5)}. \quad (37)$$

Grenham's formula reads as

$$\zeta^{\triangleleft}_{\mathfrak{G}(4),p}(s) = W_4(X, Y)|_{X=p,\, Y=p^{-s}}. \tag{38}$$

From this it is immediately clear that $\mathfrak{G}(4)$ has uniform ideal growth. Also $\alpha^{\triangleleft}_{\mathfrak{G}(4)} = 4$ and $b^{\triangleleft}_{\mathfrak{G}(4)} = 1$ can be read off. Further analysis of the numerator in (37) shows that the global zeta function $\zeta^{\triangleleft}_{\mathfrak{G}(4)}(s)$ has a natural boundary at $\mathrm{Re}(s) = 9/5$ (see [11]).

Using methods of algebraic geometry C. Voll [49] has developed a closed formula for $\zeta^{\triangleleft}_{\mathfrak{G}(n),p}(s)$ which holds for every $n \geq 2$ and every prime p. This formula shows that $\mathfrak{G}(n)$ has uniform ideal growth for every $n \geq 2$, it also confirms the conjectures from Section 5 concerning functional equations of the Euler factors. Also, for $n \geq 6$ the abscissa of convergence $\alpha^{\triangleleft}_{\mathfrak{G}(n)}$ is greater than n, and it is in general not an integer. Indeed, if $6 \leq n \leq 200$, the abscissa of convergence is an integer if and only if $n = 2N^2 + 6N + 5$ for some integer N.

D. Grenham [22] has also studied the subring zeta function of $\mathfrak{G}(n)$. We cite from [22] the following pole orders:

$$b^{<}_{\mathfrak{G}(3)} = 2, \quad b^{<}_{\mathfrak{G}(4)} = 2, \quad b^{<}_{\mathfrak{G}(5)} = 3. \tag{39}$$

The corresponding abscissas of convergence are

$$\alpha^{<}_{\mathfrak{G}(3)} = 3, \quad \alpha^{<}_{\mathfrak{G}(4)} = 4, \quad \alpha^{<}_{\mathfrak{G}(5)} = 5. \tag{40}$$

Example 3. The following example of a Lie ring played an important role in the development of the conjectures from Section 5 concerning functional equations of the Euler factors. Define

$$\mathfrak{L}_W := \langle\, z,\ w_1,\ w_2,\ x_1,\ x_2,\ y \mid [z, w_1] = x_1,\ [z, w_2] = x_2,\ [z, x_1] = y \,\rangle. \tag{41}$$

This Lie ring has nilpotency class 3 and $\mathbb{Z}$-rank 6. It was discovered and extensively studied by L. Woodward in [51]. The Lie ring $\mathfrak{L}_W$ has uniform subring and ideal growth but only the local subring counting zeta function satisfies a functional equation. We further report from [51]:

$$\alpha^{<}_{\mathfrak{L}_W} = 3,\ b^{<}_{\mathfrak{L}_W} = 4, \quad \alpha^{\triangleleft}_{\mathfrak{L}_W} = 3,\ b^{\triangleleft}_{\mathfrak{L}_W} = 1. \tag{42}$$

The global zeta function $\zeta^{<}_{\mathfrak{L}_W}(s)$ has a natural boundary at $\mathrm{Re}(s) = 17/7$ whereas $\zeta^{\triangleleft}_{\mathfrak{L}_W}(s)$ has a natural boundary at $\mathrm{Re}(s) = 7/6$.

Example 4. In [23] it is proved that the following two finitely generated nilpotent groups

$$G_1 = \left\langle \begin{array}{c} g_1, g_2, g_3, g_4, \\ z_1,\ z_2 \end{array} \;\middle|\; \begin{array}{c} [g_1, g_2] = 1,\ [g_3, g_4] = 1,\ [g_1, g_3] = z_1, \\ [g_1, g_4] = z_2,\ [g_2, g_3] = z_2,\ [g_2, g_4] = z_1^{-5} \end{array} \right\rangle, \tag{43}$$

$$G_2 = \left\langle \begin{array}{c} g_1, g_2, g_3, g_4, \\ z_1,\ z_2 \end{array} \;\middle|\; \begin{array}{c} [g_1, g_2] = 1,\ [g_3, g_4] = 1,\ [g_1, g_3] = z_1, \\ [g_1, g_4] = z_2,\ [g_2, g_3] = z_1^{-1} z_2^2,\ [g_2, g_4] = z_1^{-3} z_2 \end{array} \right\rangle \tag{44}$$

have the same profinite completion but are not isomorphic. It follows that both their zeta functions are the same. Both groups have Hirsch-length equal to 6 and are of nilpotency class 2. These groups come as special cases of an infinite series of l-tuples ($l \geq 2$) of examples of such groups arising from a number theoretic setting.

7. Variation

We have put the emphasis on counting subgroups or normal subgroups in nilpotent groups and on counting subrings or ideals in rings, however our results extend in a number of other directions.

(1) Variants of our zeta functions have been considered which count only subgroups with some added feature, for example characteristic subgroups or subgroups of a finitely generated torsion-free nilpotent group G which are isomorphic to G. Theorems 1.2 and 1.3 hold in this case and for many of these variants. In fact, there is always a p-adic formalism like in Section 2 which reduces Theorem 1.3 to Corollary 3.1 (see [27]). The paper [19] relates the zeta functions counting subgroups of G which are isomorphic to G to zeta functions defined by A. Weil for $\mathbb{Q}$-defined linear algebraic groups.

(2) The rationality result of Theorem 1.2 also holds for finitely generated nilpotent groups which are not necessarily torsion-free. In fact, the first author proved in [3] that this result extends to all finitely generated soluble groups of finite rank.

(3) In [12] it is proved that all crystallographic groups or more generally all finitely generated groups which contain an abelian subgroup of finite index have zeta functions which have a meromorphic continuation to all of $\mathbb{C}$. This is done by relating these zeta functions to zeta functions of orders in central simple $\mathbb{Q}$-algebras.

(4) The local zeta functions of the classical groups (see [14], [13]) can be expressed as p-adic cone integrals and our results apply to the corresponding Euler product.

(5) Let $g(n, c, d)$ be the number of finite nilpotent groups of size n, of nilpotency class bounded by c and generated by at most d elements. In [7] the zeta function

$$\zeta_{\mathcal{N}(c,d)}(s) := \sum_{n=1}^{\infty} g(n, c, d) n^{-s} \tag{45}$$

is shown to be expressible as the Euler product of p-adic cone integrals. Hence our results apply and give asymptotic results for the partial sums of the $g(n, c, d)$. The formalism of zeta functions has been applied successfully in [7] to solve conjecture **P**, which had appeared in connection with periodicity in trees connected with the classification problem for finite p-groups in terms of coclass.

(6) Thinking of Hilbert's basis theorem we might expect a connection between the ideal counting zeta function of a ring R and that of the polynomial ring $R[x]$ over R. This expectation is confirmed by a beautiful formula of D. Segal [45] which holds for Dedekind rings R.

(7) The formalism of zeta functions has been used to count representations of arithmetic and p-adic analytic groups in the papers [37] of B. Martin and A. Lubotzky, [30] of A. Jaikin-Zapirain and [33] of M. Larsen and A. Lubotzky.

References

[1] Denef, J., The rationality of the Poincaré series associated to the p-adic points on a variety. *Invent. Math.* **77** (1984), 1–23.

[2] Denef, J., Meuser, D., A functional equation of Igusa's local zeta function. *Amer. J. Math.* **113** (1991), 1135–1152.

[3] du Sautoy, M. P. F., Finitely generated groups, p-adic analytic groups and Poincaré series. *Ann. of Math.* **137** (1993), 639–670.

[4] du Sautoy, M. P. F., Zeta functions of groups and Lie algebras: uniformity. *Israel J. Math.* **86** (1994), 1–23.

[5] du Sautoy, M. P. F., The zeta function of $sl_2(\mathbb{Z})$. *Forum Math.* **12** (2000), 197–221.

[6] du Sautoy, M. P. F., Addendum to the paper: The zeta function of $sl_2(\mathbb{Z})$. *Forum Math.* **12** (2000), 383.

[7] du Sautoy, M. P. F., Counting p-groups and nilpotent groups. *Inst. Hautes Études Sci. Publ. Math.* **92** (2000), 63–112.

[8] du Sautoy, M. P. F., A nilpotent group and its elliptic curve: non-uniformity of local zeta functions of groups. *Israel J. Math.* **126** (2001), 269–288.

[9] du Sautoy, M. P. F., Counting subgroups in nilpotent groups and points on elliptic curves. *J. Reine Angew. Math.* **549** (2002), 1–21.

[10] du Sautoy, M. P. F., Zeta functions of groups: the quest for order versus the flight from ennui. *Groups St. Andrews 2001 in Oxford*, Vol. I, , London Math. Soc. Lecture Note Ser. 304, Cambridge University Press, Cambridge 2003, 150–189.

[11] du Sautoy, M. P. F., Natural boundaries for zeta functions of groups. Preprint.

[12] du Sautoy, M. P. F., McDermott, J. J., Smith, G. C., Zeta functions of crystallographic groups and meromorphic continuation. *Proc. London Math. Soc.* **79** (1999), 511–534.

[13] du Sautoy, M. P. F., Grunewald, F., Analytic properties of zeta functions and subgroup growth. *Ann. of Math.* **152** (2000), 793–833.

[14] du Sautoy, M. P. F., Grunewald, F., Zeta functions of classical groups and their friendly ghosts. *C. R. Acad. Sci. Paris Sér. I Math.* **327** (1998), 1–6.

[15] du Sautoy, M. P. F., Grunewald, F., Analytic properties of Euler products of Igusa type zeta functions and subgroup growth of nilpotent groups. *C. R. Acad. Sci. Paris Sér. I Math.* **329** (1999), 351–356.

[16] du Sautoy, M. P. F., Grunewald, F., Zeta functions of groups: Zeros and friendly ghosts. *Amer. J. Math.* **124** (2002), 1–48.

[17] du Sautoy, M. P. F., Grunewald, F., Natural boundaries for Euler products of Igusa zeta functions of elliptic curves. Preprint.

[18] du Sautoy, M. P. F., Loeser, F., Motivic zeta functions of infinite dimensional Lie algebras. *Selecta Math.* **10** (2004), 253–303.

[19] du Sautoy, M. P. F., Lubotzky, A., Functional equations and uniformity for local zeta-functions of nilpotent groups. *Amer. J. Math.* **118** (1996), 39–90.

[20] du Sautoy, M. P. F., Segal, D., Zeta functions of groups. In *New Horizons in pro-p Groups* (ed. by M. P. F. du Sautoy, D. Segal, A. Shalev), Progr. Math. 184, Birkhäuser, Boston 2002, 249–286.

[21] du Sautoy, M. P. F., Taylor, G., The zeta function of sl_2 and resolution of singularities. *Math. Proc. Cambridge Philos. Soc.* **132** (2002), 57–73.

[22] Grenham, D., Some topics in nilotent group theory. DPhil Thesis, Oxford, 1988.

[23] Grunewald, F., Scharlau, R., A note on finitely generated torsion-free groups of class 2. *J. Algebra* **58** (1979), 162–175.

[24] Grunewald, F., Segal, D., Some general algorithms. II: Nilpotent Groups. *Ann. of Math.* **112** (1980), 585–617.

[25] Grunewald, F., Segal, D., Nilpotent groups of Hirsch length six. *Math. Z.* **179** (1982), 219–235.

[26] Grunewald, F., Segal, D., Reflections on the classification of torsion-free nilpotent groups. In *Group theory* (Essays for Philip Hall), ed. by K. W. Gruenberg and J. E. Roseblade, Academic Press, London 1984.

[27] Grunewald, F. J., Segal, D., Smith, G. C., Subgroups of finite index in nilpotent groups. *Invent. Math.* **93** (1988), 185–223.

[28] Igusa, J., *An introduction to the theory of local zeta functions*. AMS/IP Stud. Adv. Math. 14, Amer. Math. Soc., Providence, RI, International Press, Cambrigde, MA, 2002.

[29] Ilani, I., Zeta functions relating to the group $\mathrm{SL}_2(\mathbb{Z}_p)$. *Israel J. Math.* **109** (1999), 157–172.

[30] Jaikin-Zapirain, A., Zeta function of representations of compact p-adic analytic groups. *J. Amer. Math. Soc.* **19** (2006), 91–118.

[31] Klopsch, B., Voll, C., Igusa-type functions associated to finite formed spaces and their functional equations. Preprint, 2006.

[32] Lang, S., *Algebraic Number Theory*. Addison Wesley Publishing Company, Reading, MA, 1970.

[33] Larsen, M., Lubotzky, A., Counting representations: zeta functions and rates of growth. Preprint.

[34] Lubotzky, A., Counting finite index subgroups. In *Groups '93 Galway/St. Andrews*, Vol. 2, London Math. Soc. Lecture Note Ser. 212, Cambridge University Press, Cambridge 1995, 368–404.

[35] Lubotzky, A., Subgroup growth. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 309–317.

[36] Lubotzky, A., Mann, A., and Segal, D., Finitely generated groups of polynomial subgroup growth. *Israel J. Math.* **82** (1993) 363-371.

[37] Lubotzky, A., Martin, B., Polynomial representation growth and the congruence subgroup problem. *Israel J. Math.* **144** (2004), 293–316.

[38] Narkiewicz, W., *Number Theory*. World Scientific Publishing Co., Singapore 1983.

[39] Ono, T., An integral attached to a hypersurface. *Amer. J. Math.* **90** (1968), 1224–1236.

[40] Paajanen, P. M., The normal zeta function of the free class two nilpotent group on four generators. *Geom. Dedicata* **115** (2005), 135–163.

[41] Paajanen, P. M., Zeta functions of groups and arithmetic geometry. DPhil. Thesis, Oxford, 2005.

[42] Paajanen, P. M., On the degree of polynomial subgroup growth in class 2 nilpotent groups. *Israel J. Math.*, to appear.

[43] Paajanen, P. M., Zeta functions of groups and the Segre surface. Preprint.

[44] Perlis, R., On the equation $\zeta_K(s) = \zeta_{K'}(s)$. *J. Number Theory* **9** (1977), 342–360.

[45] Segal, D., Ideals of finite index in a polynomial ring. *Quart. J. Math. Oxford* **48** (1997), 83–92.

[46] Smith, G. C., Zeta functions of torsion free finitely generated nilpotent groups. Ph.D. thesis, Manchester (UMIST), 1983.

[47] Taylor, G., Zeta functions of algebras and resolution of singularities. Ph.D. thesis, Cambridge, 2001.

[48] Voll, C., Zeta functions of groups and enumeration in Bruhat-Tits buildings. *Amer. J. Math.* **126** (2004), 1005–1032.

[49] Voll, C., Functional equations for local normal zeta functions of nilpotent groups. *Geom. Funct. Anal.* **15** (2005), 274–295.

[50] Voll, C., Normal subgroup growth in free class 2 nilpotent groups. *Math. Ann.* **332** (2005), 67–79.

[51] Woodward, L., Zeta functions of groups: computer calculations and functional equations. DPhil Thesis, Oxford, 2005.

[52] Woodward, L., Zeta functions of Lie rings archive. http://www.lack-of.org.uk/zfarchive/.

Mathematical Institute, 24–29 St Giles, Oxford OX1 3LB, UK
E-mail: dusautoy@maths.ox.ac.uk

Mathematisches Institut, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany
E-mail: fritz@math.uni-duesseldorf.de

On differential graded categories

Bernhard Keller

Abstract. Differential graded categories enhance our understanding of triangulated categories appearing in algebra and geometry. In this survey, we review their foundations and report on recent work by Drinfeld, Dugger–Shipley, Toën, Toën–Vaquié, and others.

Mathematics Subject Classification (2000). Primary 18E30; Secondary 16D90.

Keywords. Homological algebra, derived category, homotopy category, derived functor, Hochschild cohomology, K-theory, Morita theory, non-commutative algebraic geometry.

1. Introduction

1.1. Triangulated categories and dg categories. Derived categories were invented by Grothendieck–Verdier in the early sixties in order to formulate Grothendieck's duality theory for schemes, *cf.* [69]. Today, they have become an important tool in many branches of algebraic geometry, in algebraic analysis, non-commutative algebraic geometry, representation theory, mathematical physics .... In an attempt to axiomatize the properties of derived categories, Grothendieck–Verdier introduced the notion of a triangulated category. For a long time, triangulated categories were considered too poor to allow the development of more than a rudimentary theory. This vision has changed in recent years [112], [113], but the fact remains that many important constructions of derived categories cannot be performed with triangulated categories. Notably, tensor products and functor categories formed from triangulated categories are no longer triangulated. One approach to overcome these problems has been the theory of derivators initiated by Heller [63] and Grothendieck [59], *cf.* also [75], at the beginning of the nineties. Another, perhaps less formidable one is the theory of differential graded categories (= dg categories), together with its cousin, the theory of A_∞-categories.

Dg categories already appear in [86]. In the seventies, they found applications [130], [35] in the representation theory of finite-dimensional algebras. The idea to use dg categories to 'enhance' triangulated categories goes back at least to Bondal–Kapranov [21], who were motivated by the study of exceptional collections of coherent sheaves on projective varieties.

The synthesis of Koszul duality [11], [12] with Morita theory for derived categories [124] was the aim of the study of the unbounded derived category of a dg category in [76].

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

It is now well-established that invariants like K-theory, Hochschild (co-)homology and cyclic homology associated with a ring or a variety 'only depend' on its derived category. However, in most cases, the derived category (even with its triangulated structure) is not enough to compute the invariant, and the datum of a triangle equivalence between derived categories is not enough to construct an isomorphism between invariants (*cf.* Dugger–Shipley's [38] results results in Section 3.9). Differential graded categories provide the necessary structure to fill this gap. This idea was applied to K-theory by Thomason–Trobaugh [152] and to cyclic homology in [78], [80].

The most useful operation which *can be performed* on triangulated categories is the passage to a Verdier quotient. It was therefore important to lift this operation to the world of differential graded categories. This was done implicitly in [80] but explicitly, by Drinfeld, in [34].

In a certain sense, differential graded categories and differential graded functors contain too much information and the main problem in working with them consists in 'discarding what is irrelevant'. It now appears clearly that the best tool for doing this are Quillen model categories [121]: They provide a homotopy theoretic framework which allows simple, yet precise statements and rigorous but readable proofs. Building on the techniques of [34], a suitable model structure on the category of small differential graded categories was constructed in [146]. Starting from this structure, Toën has given a new approach to Morita theory for dg categories [155]. In their joint work [156], Toën and Vaquié have applied this to the construction of moduli stacks of objects in dg categories, and notably in categories of perfect complexes arising in geometry and representation theory.

Thanks to [155], [87] and to recent work by Tamarkin [148], we are perhaps getting closer to answering Drinfeld's question [34]: *What do DG categories form?*

1.2. Contents. After introducing notations and basic definitions in Section 2 we review the derived category of a dg category in Section 3. This is the first opportunity to practice the language of model categories. We present the structure theorems for algebraic triangulated categories which are compactly generated or, more generally, well-generated. We conclude with a survey of recent important work by Dugger and Shipley on topological Morita equivalence for dg categories. In Section 4, we present the homotopy categories of dg categories and of 'triangulated' dg categories following Toën's work [155]. The most important points are the description of the mapping spaces of the homotopy category via quasi-functors (Theorem 4.3), the closed monoidal structure (Theorem 4.5) and the characterization of dg categories of finite type (Theorem 4.12). We conclude with a summary of the applications to moduli problems. In the final Section 5, we present the most important invariance results for K-theory, Hochschild (co-)homology and cyclic homology. The derived Hall algebra presented in Section 5.6 is a new invariant due to Toën [153]. Its further development might lead to significant applications in representation theory.

Acknowledgments. I thank Bertrand Toën, Henning Krause, Brooke Shipley and Gonçalo Tabuada for helpful comments on previous versions of this article.

2. Definitions

2.1. Notations. Let k be a commutative ring, for example a field or the ring of integers $\mathbb{Z}$. We will write $\otimes$ for the tensor product over k. Recall that a *k-algebra* is a k-module A endowed with a k-linear associative multiplication $A \otimes_k A \to A$ admitting a two-sided unit $1 \in A$. For example, a $\mathbb{Z}$-algebra is just a (possibly non-commutative) ring. A *k-category* $\mathcal{A}$ is a 'k-algebra with several objects' in the sense of Mitchell [106]. Thus, it is the datum of a class of objects $\mathrm{obj}(\mathcal{A})$, of a k-module $\mathcal{A}(X, Y)$ for all objects X, Y of $\mathcal{A}$, and of k-linear associative composition maps

$$\mathcal{A}(Y, Z) \otimes \mathcal{A}(X, Y) \to \mathcal{A}(X, Z), \quad (f, g) \mapsto fg$$

admitting units $\mathbf{1}_X \in \mathcal{A}(X, X)$. For example, we can view k-algebras as k-categories with one object. The category $\mathrm{Mod}\, A$ of right A-modules over a k-algebra A is an example of a k-category with many objects. It is also an example of a *k-linear* category, *i.e.* a k-category which admits all finite direct sums.

A *graded k-module* is a k-module V together with a decomposition indexed by the positive and the negative integers:

$$V = \bigoplus_{p \in \mathbb{Z}} V^p.$$

The *shifted module* $V[1]$ is defined by $V[1]^p = V^{p+1}$, $p \in \mathbb{Z}$. A *morphism* $f : V \to V'$ of graded k-modules of degree n is a k-linear morphism such that $f(V^p) \subset V^{p+n}$ for all $p \in \mathbb{Z}$. The *tensor product* $V \otimes W$ of two graded k-modules V and W is the graded k-module with components

$$(V \otimes W)^n = \bigoplus_{p+q=n} V^p \otimes W^q, \quad n \in \mathbb{Z}.$$

The *tensor product* $f \otimes g$ of two maps $f : V \to V'$ and $g : W \to W'$ of graded k-modules is defined using the *Koszul sign rule*: We have

$$(f \otimes g)(v \otimes w) = (-1)^{pq} f(v) \otimes g(w)$$

if g is of degree p and v belongs to V^q. A *graded k-algebra* is a graded k-module A endowed with a multiplication morphism $A \otimes A \to A$ which is graded of degree 0, associative and admits a unit $1 \in A^0$. We identify 'ordinary' k-algebras with graded k-algebras concentrated in degree 0. We write $\mathcal{G}(k)$ for the *category of graded k-modules*.

A *differential graded* ($= dg$) *k-module* is a $\mathbb{Z}$-graded k-module V endowed with a *differential* d_V, *i.e.* a map $d_V : V \to V$ of degree 1 such that $d_V^2 = 0$. Equivalently, V is a *complex* of k-modules. The *shifted dg module* $V[1]$ is the shifted graded module endowed with the differential $-d_V$. The *tensor product* of two dg k-modules is the graded module $V \otimes W$ endowed with the differential $d_V \otimes \mathbf{1}_W + \mathbf{1}_V \otimes d_W$.

2.2. Differential graded categories. A *differential graded* or *dg category* is a k-category $\mathcal{A}$ whose morphism spaces are dg k-modules and whose compositions

$$\mathcal{A}(Y, Z) \otimes \mathcal{A}(X, Y) \to \mathcal{A}(X, Z)$$

are morphisms of dg k-modules.

For example, dg categories with one object may be identified with *dg algebras*, *i.e.* graded k-algebras endowed with a differential d such that the Leibniz rule holds:

$$d(fg) = d(f)g + (-1)^p f\, d(g)$$

for all $f \in A^p$ and all g. In particular, each ordinary k-algebra yields a dg category with one object. A typical example with several objects is obtained as follows: Let B be a k-algebra and $\mathcal{C}(B)$ the category of complexes of right B-modules

$$\cdots \longrightarrow M^p \xrightarrow{d_M} M^{p+1} \longrightarrow \cdots, \quad p \in \mathbb{Z}\,.$$

For two complexes L, M and an integer $n \in \mathbb{Z}$, we define $\mathcal{H}om(L, M)^n$ to be the k-module formed by the morphisms $f\colon L \to M$ of graded objects of degree n, *i.e.* the families $f = (f^p)$ of morphisms $f^p\colon L^p \to M^{p+n}$, $p \in \mathbb{Z}$, of B-modules. We define $\mathcal{H}om(L, M)$ to be the graded k-module with components $\mathcal{H}om(L, M)^n$ and whose differential is the commutator

$$d(f) = d_M \circ f - (-1)^n f \circ d_L.$$

The *dg category* $\mathcal{C}_{\mathrm{dg}}(B)$ has as objects all complexes and its morphisms are defined by

$$\mathcal{C}_{\mathrm{dg}}(B)(L, M) = \mathcal{H}om(L, M).$$

The composition is the composition of graded maps.

Let $\mathcal{A}$ be a dg category. The *opposite dg category* $\mathcal{A}^{\mathrm{op}}$ has the same objects as $\mathcal{A}$ and its morphisms are defined by

$$\mathcal{A}^{\mathrm{op}}(X, Y) = \mathcal{A}(Y, X);$$

the composition of $f \in \mathcal{A}^{\mathrm{op}}(Y, X)^p$ with $g \in \mathcal{A}^{\mathrm{op}}(Z, Y)^q$ is given by $(-1)^{pq} gf$. The *category* $Z^0(\mathcal{A})$ has the same objects as $\mathcal{A}$ and its morphisms are defined by

$$(Z^0\mathcal{A})(X, Y) = Z^0(\mathcal{A}(X, Y)),$$

where Z^0 is the kernel of $d\colon \mathcal{A}(X, Y)^0 \to \mathcal{A}(X, Y)^1$. The *category* $H^0(\mathcal{A})$ has the same objects as $\mathcal{A}$ and its morphisms are defined by

$$(H^0(\mathcal{A}))(X, Y) = H^0(\mathcal{A}(X, Y)),$$

where H^0 denotes the 0th homology of the complex $\mathcal{A}(X, Y)$. For example, if B is a k-algebra, we have an isomorphism of categories

$$Z^0(\mathcal{C}_{\mathrm{dg}}(B)) = \mathcal{C}(B)$$

and an isomorphism of categories

$$H^0(\mathcal{C}_{\mathrm{dg}}(B)) = \mathcal{H}(B),$$

where $\mathcal{H}(B)$ denotes the *category of complexes up to homotopy*, *i.e.* the category whose objects are the complexes and whose morphisms are the morphisms of complexes modulo the morphisms f homotopic to zero, *i.e.* such that $f = d(g)$ for some $g \in \mathcal{H}om(L, M)^{-1}$. The *homology category* $H^*(\mathcal{A})$ is the graded category with the same objects as $\mathcal{A}$ and morphisms spaces $H^*\mathcal{A}(X, Y)$.

2.3. The category of dg categories. Let $\mathcal{A}$ and $\mathcal{A}'$ be dg categories. A *dg functor* $F\colon \mathcal{A} \to \mathcal{A}'$ is given by a map $F\colon \mathrm{obj}(\mathcal{A}) \to \mathrm{obj}(\mathcal{A}')$ and by morphisms of dg k-modules

$$F(X, Y)\colon \mathcal{A}(X, Y) \to \mathcal{A}(FX, FY), \quad X, Y \in \mathrm{obj}(\mathcal{A}),$$

compatible with the composition and the units. The *category of small dg categories* dgcat_k has the small dg categories as objects and the dg functors as morphisms. Note that it has an initial object, the empty dg category $\emptyset$, and a final object, the dg category with one object whose endomorphism ring is the zero ring. The *tensor product* $\mathcal{A} \otimes \mathcal{B}$ of two dg categories has the class of objects $\mathrm{obj}(\mathcal{A}) \times \mathrm{obj}(\mathcal{B})$ and the morphism spaces

$$(\mathcal{A} \otimes \mathcal{B})((X, Y), (X', Y')) = \mathcal{A}(X, X') \otimes \mathcal{B}(Y, Y')$$

with the natural compositions and units.

For two dg functors $F, G\colon \mathcal{A} \to \mathcal{B}$, the *complex of graded morphisms* $\mathcal{H}om(F, G)$ has as its nth component the module formed by the families of morphisms

$$\phi_X \in \mathcal{B}(FX, GX)^n$$

such that $(Gf)(\phi_X) = (\phi_Y)(Ff)$ for all $f \in \mathcal{A}(X, Y)$, $X, Y \in \mathcal{A}$. The differential is induced by that of $\mathcal{B}(FX, GX)$. The set of *morphisms* $F \to G$ is by definition in bijection with $Z^0\mathcal{H}om(F, G)$.

Endowed with the tensor product, the category dgcat_k becomes a symmetric tensor category which admits an internal Hom-functor, *i.e.*

$$\mathrm{Hom}(\mathcal{A} \otimes \mathcal{B}, \mathcal{C}) = \mathrm{Hom}(\mathcal{A}, \mathcal{H}om(\mathcal{B}, \mathcal{C})),$$

for $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathrm{dgcat}_k$, where $\mathcal{H}om(\mathcal{B}, \mathcal{C})$ has the dg functors as objects and the morphism space $\mathcal{H}om(F, G)$ for two dg functors F and G. The unit object is the dg category associated with the k-algebra k.

A *quasi-equivalence* is a dg functor $F\colon \mathcal{A} \to \mathcal{A}'$ such that

1) $F(X, Y)$ is a quasi-isomorphism for all objects X, Y of $\mathcal{A}$, and

2) the induced functor $H^0(F)\colon H^0(\mathcal{A}) \to H^0(\mathcal{A}')$ is an equivalence.

Note that neither the tensor product nor the internal Hom-functor respect the quasi-equivalences, a source of technical difficulties.

3. The derived category of a dg category

3.1. Dg modules. Let $\mathcal{A}$ be a small dg category. A *left dg $\mathcal{A}$-module* is a dg functor

$$L\colon \mathcal{A} \to \mathcal{C}_{\mathrm{dg}}(k)$$

and a *right dg $\mathcal{A}$-module* a dg functor

$$M\colon \mathcal{A}^{\mathrm{op}} \to \mathcal{C}_{\mathrm{dg}}(k).$$

Equivalently, a right dg $\mathcal{A}$-module M is given by complexes $M(X)$ of k-modules, for each $X \in \mathrm{obj}(\mathcal{A})$, and by morphisms of complexes

$$M(Y) \otimes \mathcal{A}(X, Y) \to M(X)$$

compatible with compositions and units. The *homology* $H^*(M)$ of a dg module M is the induced functor

$$H^*(\mathcal{A}) \to \mathcal{G}(k), \quad X \mapsto H^*(M(X))$$

with values in the category $\mathcal{G}(k)$ of graded k-modules (*cf.* 2.1). For each object X of $\mathcal{A}$, we have the right module *represented by* X

$$X^\wedge = \mathcal{A}(?, X).$$

The *category of dg modules* $\mathcal{C}(\mathcal{A})$ has as objects the dg $\mathcal{A}$-modules and as morphisms $L \to M$ the morphisms of dg functors (*cf.* 2.3). Note that $\mathcal{C}(\mathcal{A})$ is an abelian category and that a morphism $L \to M$ is an epimorphism (respectively a monomorphism) iff it induces surjections (respectively injections) in each component of $L(X) \to M(X)$ for each object X of $\mathcal{A}$. A morphism $f\colon L \to M$ is a *quasi-isomorphism* if it induces an isomorphism in homology.

We have $\mathcal{C}(\mathcal{A}) = Z^0(\mathcal{C}_{\mathrm{dg}}(\mathcal{A}))$, where, in the notations of 2.3, the dg category $\mathcal{C}_{\mathrm{dg}}(\mathcal{A})$ is defined by

$$\mathcal{C}_{\mathrm{dg}}(\mathcal{A}) = \mathcal{H}om(\mathcal{A}^{\mathrm{op}}, \mathcal{C}_{\mathrm{dg}}(k)).$$

We write $\mathcal{H}om(L, M)$ for the complex of morphisms from L to M in $\mathcal{C}_{\mathrm{dg}}(\mathcal{A})$. For each $X \in \mathcal{A}$, we have a natural isomorphism

$$\mathcal{H}om(X^\wedge, M) \xrightarrow{\sim} M(X). \tag{1}$$

The *category up to homotopy of dg $\mathcal{A}$-modules* is

$$\mathcal{H}(\mathcal{A}) = H^0(\mathcal{C}_{\mathrm{dg}}(\mathcal{A})).$$

The isomorphism (1) yields isomorphisms

$$\mathcal{H}(\mathcal{A})(X^\wedge, M[n]) \xrightarrow{\sim} H^n(\mathcal{H}om(X^\wedge, M)) \xrightarrow{\sim} H^n M(X), \tag{2}$$

where $n \in \mathbb{Z}$ and $M[n]$ is the *shifted dg module* $Y \mapsto M(Y)[n]$.

If $\mathcal{A}$ is the dg category with one object associated with a k-algebra B, then a dg $\mathcal{A}$-module is the same as a complex of B-modules. More precisely, we have $\mathcal{C}(\mathcal{A}) = \mathcal{C}(B)$, $\mathcal{C}_{\mathrm{dg}}(\mathcal{A}) = \mathcal{C}_{\mathrm{dg}}(B)$ and $\mathcal{H}(\mathcal{A}) = \mathcal{H}(B)$. In this case, if X is the unique object of $\mathcal{A}$, the dg module $X^\wedge$ is the complex formed by the free right B-module of rank one concentrated in degree 0.

3.2. The derived category, resolutions. The *derived category* $\mathcal{D}(\mathcal{A})$ is the localization of the category $\mathcal{C}(\mathcal{A})$ with respect to the class of quasi-isomorphisms. Thus, its objects are the dg modules and its morphisms are obtained from morphisms of dg modules by formally inverting [53] all quasi-isomorphisms. The projection functor $\mathcal{C}(\mathcal{A}) \to \mathcal{D}(\mathcal{A})$ induces a functor $\mathcal{H}(\mathcal{A}) \to \mathcal{D}(\mathcal{A})$ and the derived category could equivalently be defined as the localization of $\mathcal{H}(\mathcal{A})$ with respect to the class of all quasi-isomorphisms. Note that from this definition, it is not clear that the morphism classes of $\mathcal{D}(\mathcal{A})$ are sets or that $\mathcal{D}(\mathcal{A})$ is an additive category.

Call a dg module P *cofibrant* if, for every surjective quasi-isomorphism $L \to M$, every morphism $P \to M$ factors through L. For example, for an object X of $\mathcal{A}$, the dg module $X^\wedge$ is cofibrant. Call a dg module I *fibrant* if, for every injective quasi-isomorphism $L \to M$, every morphism $L \to I$ extends to M. For example, if E is an injective cogenerator of the category of k-modules and X an object of $\mathcal{A}$, the dg module $\mathcal{H}om(\mathcal{A}(X, ?), E)$ is fibrant.

Proposition 3.1. a) *For each dg module M, there is a quasi-isomorphism $\mathbf{p}M \to M$ with cofibrant $\mathbf{p}M$ and a quasi-isomorphism $M \to \mathbf{i}M$ with fibrant $\mathbf{i}M$.*

b) *The projection functor $\mathcal{H}(\mathcal{A}) \to \mathcal{D}(\mathcal{A})$ admits a fully faithful left adjoint given by $M \mapsto \mathbf{p}M$ and a fully faithful right adjoint given by $M \mapsto \mathbf{i}M$.*

One can construct $\mathbf{p}M$ and $\mathbf{i}M$ explicitly, as first done in [5] (*cf.* also [76]). We call $\mathbf{p}M \to M$ a *cofibrant resolution* and $M \to \mathbf{i}M$ a *fibrant resolution* of M. According to b), these resolutions are functorial in the category up to homotopy $\mathcal{H}(\mathcal{A})$ and we can compute morphisms in $\mathcal{D}(\mathcal{A})$ via

$$\mathcal{H}(\mathcal{A})(\mathbf{p}L, M) = \mathcal{D}(\mathcal{A})(L, M) = \mathcal{H}(\mathcal{A})(L, \mathbf{i}M).$$

In particular, for an object X of $\mathcal{A}$ and a dg module M, the isomorphisms (2) yield

$$\mathcal{D}(\mathcal{A})(X^\wedge, M[n]) \xrightarrow{\sim} \mathcal{H}(\mathcal{A})(X^\wedge, M[n]) \xrightarrow{\sim} H^n M(X) \tag{3}$$

since $X^\wedge$ is cofibrant. The embedding $\mathcal{D}(\mathcal{A}) \to \mathcal{H}(\mathcal{A})$ provided by $\mathbf{p}$ also shows that the derived category is additive.

If $\mathcal{A}$ is associated with a k-algebra B and M is a right B-module considered as a complex concentrated in degree 0, then $\mathbf{p}M \to M$ is a projective resolution of M and $M \to \mathbf{i}M$ an injective resolution. The proposition is best understood in the language of Quillen model categories [121]. We refer to [45] for a highly readable introduction and to [66], [65] for in-depth treatments. The proposition results from the following theorem, proved using the techniques of [66, 2.3].

Theorem 3.2. *The category $\mathcal{C}(\mathcal{A})$ admits two structures of Quillen model category whose weak equivalences are the quasi-isomorphisms:*

1) *The* projective *structure, whose fibrations are the epimorphisms. For this structure, each object is fibrant and an object is cofibrant iff it is a cofibrant dg module.*
2) *The* injective *structure, whose cofibrations are the monomorphisms. For this structure, each object is cofibrant and an object is fibrant iff it is a fibrant dg module.*

For both structures, two morphisms are homotopic iff they become equal in the category up to homotopy $\mathcal{H}(\mathcal{A})$.

3.3. Exact categories, Frobenius categories. Recall that an *exact category* in the sense of Quillen [120] is an additive category $\mathcal{E}$ endowed with a distinguished class of sequences

$$0 \longrightarrow A \xrightarrow{i} B \xrightarrow{p} C \longrightarrow 0\ ,$$

where i is a kernel of p and p a cokernel of i. We will state the axioms these sequences have to satisfy using the terminology of [52]: The morphisms p are called deflations, the morphisms i inflations and the pairs (i, p) conflations. The axioms are:

Ex0 The identity morphism of the zero object is a deflation.

Ex1 The composition of two deflations is a deflation.

Ex2 Deflations are stable under base change.

Ex2′ Inflations are stable under cobase change.

As shown in [74], these axioms are equivalent to Quillen's and they imply that if $\mathcal{E}$ is small, then there is a fully faithful functor from $\mathcal{E}$ into an abelian category $\mathcal{E}'$ whose image is an additive subcategory closed under extensions and such that a sequence of $\mathcal{E}$ is a conflation iff its image is a short exact sequence of $\mathcal{E}'$. Conversely, one easily checks that an extension closed full additive subcategory $\mathcal{E}$ of an abelian category $\mathcal{E}'$ endowed with all conflations which become exact sequences in $\mathcal{E}'$ is always exact. The fundamental notions and constructions of homological algebra, and in particular the construction of the derived category, naturally extend from abelian to exact categories, *cf.* [107] and [77].

A *Frobenius category* is an exact category $\mathcal{E}$ which has enough injectives and enough projectives and where the class of projectives coincides with the class of injectives. In this case, the *stable category* $\underline{\mathcal{E}}$ obtained by dividing $\mathcal{E}$ by the ideal of morphisms factoring through a projective-injective carries a canonical structure of triangulated category, *cf.* [62], [60], [85], [57]. We write $\bar{f}$ for the image in $\underline{\mathcal{E}}$ of a morphism f of $\mathcal{E}$. The suspension functor S of $\underline{\mathcal{E}}$ is obtained by choosing a conflation

$$0 \longrightarrow A \longrightarrow IA \longrightarrow SA \longrightarrow 0$$

for each object A. Each triangle is isomorphic to a standard triangle $(\bar{\imath}, \bar{p}, \bar{e})$ obtained by embedding a conflation (i, p) into a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{i} & B & \xrightarrow{p} & C & \longrightarrow & 0 \\ & & \mathbf{1}\downarrow & & \downarrow & & \downarrow e & & \\ 0 & \longrightarrow & A & \longrightarrow & IA & \longrightarrow & SA & \longrightarrow & 0. \end{array}$$

3.4. Triangulated structure. Let $\mathcal{A}$ be a small dg category. Define a sequence

$$0 \longrightarrow L \xrightarrow{i} M \xrightarrow{p} N \longrightarrow 0$$

of $\mathcal{C}(\mathcal{A})$ to be a *conflation* if there is a morphism $r \in \mathcal{H}om(M, L)^0$ such that $ri = \mathbf{1}_L$ or, equivalently, a morphism $s \in \mathcal{H}om(N, M)$ such that $ps = \mathbf{1}_N$.

Lemma 3.3. a) *Endowed with these conflations, $\mathcal{C}(\mathcal{A})$ becomes a Frobenius category. The resulting stable category is canonically isomorphic to $\mathcal{H}(\mathcal{A})$. The suspension functor is induced by the shift $M \mapsto M[1]$.*

b) *Endowed with the suspension induced by that of $\mathcal{H}(\mathcal{A})$ and the triangles isomorphic to images of triangles of $\mathcal{H}(\mathcal{A})$ the derived category $\mathcal{D}(\mathcal{A})$ becomes a triangulated category. Each short exact sequence of complexes yields a canonical triangle.*

3.5. Compact objects, Brown representability. Let $\mathcal{T}$ be a triangulated category admitting arbitrary coproducts. Since the adjoint of a triangle functor is a triangle functor [85], the coproduct of triangles is then automatically a triangle. Moreover, $\mathcal{T}$ is *idempotent complete* [18], *i.e.* each idempotent endomorphism of an object of $\mathcal{T}$ is the composition of a section with a retraction. An object C of $\mathcal{T}$ is *compact* if the functor $\mathcal{T}(C, ?)$ commutes with arbitrary coproducts, *i.e.* for each family (X_i) of objects of $\mathcal{T}$, the canonical morphism

$$\coprod \mathcal{T}(C, X_i) \to \mathcal{T}\Big(C, \coprod X_i\Big)$$

is invertible. The triangulated category $\mathcal{T}$ is *compactly generated* if there is a set $\mathcal{G}$ of compact objects G such that an object X of $\mathcal{T}$ vanishes iff we have $\mathcal{T}(G, X) = 0$ for each $G \in \mathcal{G}$.

Theorem 3.4 (Characterization of compact objects [152], [108]). *An object of $\mathcal{T}$ is compact iff it is a direct factor of an iterated extension of copies of objects of $\mathcal{G}$ shifted in both directions.*

Theorem 3.5 (Brown representability [25], [1], [109]). *If $\mathcal{T}$ is compactly generated, a cohomological functor $F\colon \mathcal{T}^{\mathrm{op}} \to \mathrm{Mod}\,\mathbb{Z}$ is representable iff it takes coproducts of $\mathcal{T}$ to products of* $\mathrm{Mod}\,\mathbb{Z}$.

A set of objects $\mathcal{G}$ *symmetrically generates* $\mathcal{T}$ [95] if we have

1) an object X of $\mathcal{T}$ vanishes iff $\mathcal{T}(G, X) = 0$ for each $G \in \mathcal{G}$, and
2) there is a set of objects $\mathcal{G}'$ such that a morphism $f: X \to Y$ of $\mathcal{T}$ induces surjections $\mathcal{T}(G, X) \to \mathcal{T}(G, Y)$ for all $G \in \mathcal{G}$ iff it induces injections $\mathcal{T}(Y, G') \to \mathcal{T}(X, G')$ for all $G' \in \mathcal{G}'$.

If $\mathcal{G}$ compactly generates $\mathcal{T}$, then we can take for $\mathcal{G}'$ the set of objects G' defined by

$$\mathcal{T}(?, G') = \mathrm{Hom}_k(\mathcal{T}(G, ?), E), \quad G \in \mathcal{G}$$

where E is an injective cogenerator of the category of k-modules. Thus, in this case, $\mathcal{G}$ also symmetrically generates $\mathcal{T}$.

Theorem 3.6 (Brown representability for the dual [111], [95]). *If $\mathcal{T}$ is symmetrically generated, a homological functor $F: \mathcal{T} \to \mathrm{Mod}\,\mathbb{Z}$ is corepresentable iff it commutes with products.*

Let $\mathcal{A}$ be a small dg category. The derived category $\mathcal{D}(\mathcal{A})$ admits arbitrary coproducts and these are induced by coproducts of modules. Thanks to the isomorphisms

$$\mathcal{D}(\mathcal{A})(X^\wedge[n], M) \xrightarrow{\sim} H^{-n}M(X) \tag{4}$$

obtained from (3), each dg module $X^\wedge[n]$, where X is an object of $\mathcal{A}$ and n an integer, is compact. The isomorphism (4) also shows that a dg module M vanishes in $\mathcal{D}(\mathcal{A})$ iff each morphism $X^\wedge[n] \to M$ vanishes. Thus the set $\mathcal{G}$ formed by the $X^\wedge[n]$, $X \in \mathcal{A}$, $n \in \mathbb{Z}$, is a set of compact generators for $\mathcal{D}(\mathcal{A})$. The *triangulated category* $\mathrm{tria}(\mathcal{A})$ *associated with* $\mathcal{A}$ is the closure in $\mathcal{D}(\mathcal{A})$ of the set of representable functors $X^\wedge$, $X \in \mathcal{A}$, under shifts in both directions and extensions. The *category of perfect objects* $\mathrm{per}(\mathcal{A})$ the closure of $\mathrm{tria}(\mathcal{A})$ under passage to direct factors in $\mathcal{D}(\mathcal{A})$. The above theorems yield the

Corollary 3.7. *An object of $\mathcal{D}(\mathcal{A})$ is compact iff it lies in* $\mathrm{per}(\mathcal{A})$. *A cohomological functor $\mathcal{D}(\mathcal{A})^{\mathrm{op}} \to \mathrm{Mod}\,k$ is representable iff it takes coproducts of $\mathcal{D}(\mathcal{A})$ to products of* $\mathrm{Mod}\,k$. *A homological functor $\mathcal{D}(\mathcal{A}) \to \mathrm{Mod}\,k$ is corepresentable iff it commutes with products.*

3.6. Algebraic triangulated categories. Let $\mathcal{T}$ be a k-linear triangulated category. We say that $\mathcal{T}$ is *algebraic* if it is triangle equivalent to $\underline{\mathcal{E}}$ for some k-linear Frobenius category $\mathcal{T}$. It is easy to see that each subcategory of an algebraic triangulated category is algebraic. We will see below that each Verdier localization of an algebraic triangulated category is algebraic (if we neglect a set-theoretic problem). Moreover, categories of complexes up to homotopy are algebraic, by 3.4. Therefore, 'all' triangulated categories occurring in algebra and geometry are algebraic. Non algebraic triangulated categories appear naturally in topology (*cf.* also Section 3.9): For instance, in the homotopy category of 2-local spectra, the identity morphism of the

cone over twice the identity of the sphere spectrum is of order four, but in each algebraic triangulated category, the identity of the cone on twice the identity of an object is of order two at most. A general method to prove that a triangulated category obtained from a suitable stable Quillen model category is not algebraic is to show that its [123] Hom-functor enriched in spectra does not factor through the canonical functor from the derived category of abelian groups to the homotopy category of spectra.

We wish to show that 'all' algebraic triangulated categories can be described by dg categories. Let $\mathcal{T}$ be a triangulated category and $\mathcal{G}$ a full subcategory. We make $\mathcal{G}$ into a graded category $\mathcal{G}_{gr}$ by defining

$$\mathcal{G}_{gr}(G, G') = \bigoplus_{n\in\mathbb{Z}} \mathcal{T}(G, S^n G').$$

We obtain a natural functor $\overline{F}$ from $\mathcal{T}$ to the category of graded $\mathcal{G}_{gr}$-modules by sending an object Y of $\mathcal{T}$ to the $\mathcal{G}_{gr}$-module

$$X \mapsto \bigoplus_{n\in\mathbb{Z}} \mathcal{T}(X, S^n Y)$$

Theorem 3.8 ([76]). *Suppose that $\mathcal{T}$ is algebraic. Then there is a dg category $\mathcal{A}$ such that $H^*(\mathcal{A})$ is isomorphic to $\mathcal{G}_{gr}$ and a triangle functor*

$$F\colon \mathcal{T} \to \mathcal{D}(\mathcal{A})$$

such that the composition $H^* \circ F$ is isomorphic to $\overline{F}$. Moreover,

a) *F induces an equivalence from $\mathcal{T}$ to* $\mathrm{tria}(\mathcal{A})$ *iff $\mathcal{T}$ coincides with its smallest full triangulated subcategory containing $\mathcal{G}$;*

b) *F induces an equivalence from $\mathcal{T}$ to* $\mathrm{per}(\mathcal{A})$ *iff $\mathcal{T}$ is idempotent complete* (cf. *Section* 3.5) *and equals the closure of $\mathcal{G}$ under shifts in both directions, extensions and passage to direct factors;*

c) *F is an equivalence $\mathcal{T} \xrightarrow{\sim} \mathcal{D}(\mathcal{A})$ iff $\mathcal{T}$ admits arbitrary coproducts and the objects of $\mathcal{G}$ form a set of compact generators for $\mathcal{T}$.*

Examples arise from commutative and non-commutative geometry: A. Bondal and M. Van den Bergh show in [19] that if X is a quasi-compact quasi-separated scheme, then the (unbounded) derived category $\mathcal{T} = \mathcal{D}_{qc}(X)$ of complexes of $\mathcal{O}_X$-modules with quasi-coherent homology admits a single compact generator G and that moreover, $\mathrm{Hom}(G, G[n])$ vanishes except for finitely many n. Thus $\mathcal{T}$ is equivalent to the derived category of a dg category with one object whose endomorphism ring has bounded homology.

R. Rouquier shows in [131] (*cf.* also [96]) that if X is a quasi-projective scheme over a perfect field k, then the derived category of coherent sheaves over X admits a generator as a triangulated category (as in part b) and, surprisingly, that it is even of 'finite dimension' as a triangulated category: each object occurs as a direct factor of

an object which admits a 'resolution' of bounded length by finite sums of shifts of the generator. Thus, the bounded derived category of coherent sheaves is equivalent to $\mathrm{per}(\mathcal{A})$ for a dg category with one object whose endomorphism ring satisfies a strong regularity condition.

In [17], J. Block describes the bounded derived category of complexes of sheaves with coherent homology on a complex manifold X as the category $H^0(\mathcal{A})$ associated with a dg category constructed from the Dolbeault dg algebra $(A^{0,\bullet}(X), \bar{\partial})$. This can be seen as an instance of a), where, for $\mathcal{G}$, we can take for example the category of coherent sheaves (*i.e.* complexes concentrated in degree 0). Note however that the term 'perfect derived category' is used with a different meaning in [17].

In the independently obtained [40], W. Dwyer and J. Greenlees give elegant descriptions via dg endomorphism rings of categories of complete, respectively torsion, modules. Their results are applied in a unifying study of duality phenomena in algebra and topology in [41].

One of the original motivations for the theorem was D. Happel's description [60], [61] of the bounded derived category of a finite-dimensional associative algebra of finite global dimension as the stable category of a certain Frobenius category. This in turn was inspired by Bernstein–Gelfand–Gelfand's [16] and Beilinson's [9] descriptions of the derived category of coherent sheaves on projective space.

A vast generalization of the theorem to non-additive contexts [137] is due to S. Schwede and B. Shipley [139], *cf.* also Section 3.9 below.

3.7. Well-generated algebraic triangulated categories. A triangulated category $\mathcal{T}$ is *well-generated* [112], [94] if it admits arbitrary coproducts and a *good set of generators* $\mathcal{G}$, *i.e.* $\mathcal{G}$ is stable under shifts in both directions and satisfies

1) an object X of $\mathcal{T}$ vanishes iff $\mathcal{T}(G, X) = 0$ for each $G \in \mathcal{G}$,
2) there is a cardinal α such that each $G \in \mathcal{G}$ is α-*compact*, *i.e.* for each family of objects X_i, $i \in I$, of $\mathcal{T}$, each morphism
$$X \to \bigoplus_{i \in I} X_i$$
factors through a subsum $\bigoplus_{i \in J} X_i$ for some subset J of I of cardinality strictly smaller than α,
3) for each family of morphisms $f_i \colon X_i \to Y_i$, $i \in I$, of $\mathcal{T}$ which induces surjections
$$\mathcal{T}(G, X_i) \to \mathcal{T}(G, Y_i)$$
for all $G \in \mathcal{G}$ and all $i \in I$, the sum of the f_i induces surjections
$$\mathcal{T}(G, \bigoplus X_i) \to \mathcal{T}(G, \bigoplus Y_i)$$
for all $G \in \mathcal{G}$.

Clearly each compactly generated triangulated category is well-generated. A. Neeman proves in [112] that the Brown representability theorem holds for well-generated triangulated categories. This is one of the main reasons for studying them. Another important result of [112] is that if $\mathcal{T}$ is well-generated and $\mathcal{S} \to \mathcal{T}$ is a localization (*i.e.* a fully faithful triangle functor admitting a left adjoint whose kernel is generated by a *set* of objects) then $\mathcal{S}$ is well-generated. Thus each localization of a compactly generated triangulated category is well-generated and in particular, so is each localization of the derived category of a small dg category.

Here is another class of examples: Let $\mathcal{B}$ be a Grothendieck abelian category, *e.g.* the category of modules on a ringed space. Then, by the Popescu–Gabriel theorem [118], [101], $\mathcal{B}$ is the localization of the category of Mod A of A-modules over some ring A. One can deduce from this that the unbounded derived category of the abelian category $\mathcal{B}$ (*cf.* [51], [151], [71]) is a localization of $\mathcal{D}(A)$ and thus is well-generated.

Theorem 3.9 ([119]). *Let $\mathcal{T}$ be an algebraic triangulated category. Then $\mathcal{T}$ is well-generated iff it is a localization of $\mathcal{D}(\mathcal{A})$ for some small dg category $\mathcal{A}$. Moreover, if $\mathcal{T}$ is well-generated and $\mathcal{U} \subset \mathcal{T}$ a full small subcategory such that, for each $X \in \mathcal{T}$, we have*

$$X = 0 \Leftrightarrow \mathcal{T}(U, S^n X) = 0 \text{ for all } n \in \mathbb{Z} \text{ and } U \in \mathcal{U},$$

then there is an associated localization $\mathcal{T} \to \mathcal{D}(\mathcal{A})$ for some small dg category $\mathcal{A}$ with $H^*(\mathcal{A}) = \mathcal{U}_{gr}$.

3.8. Morita equivalence. Let $\mathcal{A}$ and $\mathcal{B}$ be small dg categories. Let X be an $\mathcal{A}$-$\mathcal{B}$-*bimodule*, *i.e.* a dg $\mathcal{A}^{\mathrm{op}} \otimes \mathcal{B}$-module X. Thus X is given by complexes $X(B, A)$, for all A in $\mathcal{A}$ and B in $\mathcal{B}$, and morphisms of complexes

$$\mathcal{B}(A, A') \otimes X(B, A) \otimes \mathcal{A}(B', B) \to X(B', A').$$

For each dg $\mathcal{B}$-module M, we obtain a dg $\mathcal{A}$-module

$$GM = \mathcal{H}om(X, M)\colon A \mapsto \mathcal{H}om(X(?, A), M).$$

The functor $G\colon \mathcal{C}(\mathcal{B}) \to \mathcal{C}(\mathcal{A})$ admits a left adjoint $F\colon L \mapsto L \otimes_{\mathcal{A}} X$. These functors do not respect quasi-isomorphisms in general, but they form a Quillen adjunction (*cf.* Section 3.9) and their derived functors

$$\mathbf{L}F\colon L \mapsto F(\mathbf{p}L) \quad \text{and} \quad \mathbf{R}G\colon M \mapsto G(\mathbf{i}M)$$

form an adjoint pair of functors between $\mathcal{D}(\mathcal{A})$ and $\mathcal{D}(\mathcal{B})$.

Lemma 3.10 ([76]). *The functor $\mathbf{L}F\colon \mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$ is an equivalence if and only if*

a) *the dg $\mathcal{B}$-module $X(?, A)$ is perfect for all A in $\mathcal{A}$,*

b) *the morphism*

$$\mathcal{A}(A, A') \to \mathcal{H}om(X(?, A), X(?, A'))$$

is a quasi-isomorphism for all A, A' in $\mathcal{A}$, and

c) *the dg $\mathcal{B}$-modules $X(?, A)$, $A \in \mathcal{A}$, form a set of (compact) generators for $\mathcal{D}(\mathcal{B})$.*

For example, if $E\colon \mathcal{A} \to \mathcal{B}$ is a dg functor, then $X(B, A) = \mathcal{B}(B, E(A))$ defines a dg bimodule so that the above functor G is the restriction along E. Then the lemma shows that $\mathbf{R}G$ is an equivalence iff E is a quasi-equivalence. We loosely refer to the functor $\mathbf{L}F$ associated with a dg $\mathcal{A}$-$\mathcal{B}$-bimodule as a *tensor functor*.

Theorem 3.11 ([76]). *The following are equivalent:*

1) *There is an equivalence $\mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$ given by a composition of tensor functors and their inverses.*

2) *There is a dg subcategory $\mathcal{G}$ of $\mathcal{C}(\mathcal{B})$ formed by cofibrant dg modules such that the objects of $\mathcal{G}$ form a set of compact generators for $\mathcal{D}(\mathcal{B})$ and there is a chain of quasi-equivalences*

$$\mathcal{A} \leftarrow \mathcal{A}' \to \cdots \leftarrow \mathcal{G}' \to \mathcal{G}$$

linking $\mathcal{A}$ to $\mathcal{G}$.

We say that $\mathcal{A}$ and $\mathcal{B}$ are *dg Morita equivalent* if the conditions of the theorem are satisfied. In this case, there is of course a triangle equivalence $\mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$. In general, the existence of such a triangle equivalence is not sufficient for $\mathcal{A}$ and $\mathcal{B}$ to be dg Morita equivalent, *cf.* Section 3.9. The following theorem is therefore remarkable:

Theorem 3.12 (Rickard [124]). *Suppose that $\mathcal{A}$ and $\mathcal{B}$ have their homology concentrated in degree* 0*. Then the following are equivalent:*

1) *$\mathcal{A}$ and $\mathcal{B}$ are dg Morita equivalent.*

2) *There is a triangle equivalence $\mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$.*

3) *There is a full subcategory $\mathcal{T}$ of $\mathcal{D}(\mathcal{B})$ such that*

   a) *the objects of $\mathcal{T}$ form a set of compact generators of $\mathcal{D}(\mathcal{B})$,*

   b) *we have $\mathcal{D}(\mathcal{B})(T, T'[n]) = 0$ for all $n \neq 0$ and all T, T' of $\mathcal{T}$,*

   c) *there is an equivalence $H^0(\mathcal{A}) \xrightarrow{\sim} \mathcal{T}$.*

We refer to [76] or [36] for this form of the theorem. A subcategory $\mathcal{T}$ satisfying a) and b) in 3) is called a *tilting subcategory*, a concept which generalizes that of a tilting module. We refer to [122], [3] for the theory of tilting, from which this theorem arose and which provides huge classes of examples from the representation theory of finite-dimensional algebras and finite groups as well as from algebraic geometry, *cf.* also the appendix to [113] and [88], [125], [132].

3.9. Topological Morita equivalence. In recent years, Morita theory has been vastly generalized from algebraic triangulated categories to stable model categories in work due to S. Schwede and B. Shipley. This is based on the category of symmetric spectra as constructed in [67] (*cf.* [46] for a different construction of a symmetric monoidal model category for the category of spectra). We refer to [136] for an excellent exposition of these far-reaching results and their surprising applications in homotopy theory. In work by D. Dugger and B. Shipley [38], *cf.* also [37], [39], this 'topological Morita theory' has been applied to dg categories. We briefly describe their results and refer to [141] for a highly readable, more detailed survey.

The main idea is to replace the monoidal base category, the derived category of abelian groups $\mathcal{D}(\mathbb{Z})$, by a more fundamental category: the 'derived category of the category of sets', *i.e.* the homotopy category of spectra. To preserve higher homotopical information, one must not, of course, work at the level of derived categories but has to introduce model categories. So instead of considering $\mathcal{D}(\mathbb{Z})$, one considers its model category $\mathcal{C}(\mathbb{Z})$ of complexes of abelian groups and replaces it by a convenient model of the category of spectra: the category of symmetric spectra, which one might imagine as 'complexes of abelian groups up to homotopy'. We refer to [67] or [136] for the precise definition. As shown in [67], symmetric spectra form a *symmetric* monoidal category which carries a compatible Quillen model structure and whose homotopy category is equivalent to the homotopy category of spectra of Bousfield and Friedlander [24]. The tensor product is the *smash product* $\wedge$ and the unit object is the *sphere spectrum* $\mathbb{S}$. The unit object is cofibrant and the smash product induces a monoidal structure on the homotopy category of symmetric spectra. The *Eilenberg–MacLane functor* H is a lax monoidal functor from the category of complexes $\mathcal{C}(k)$ to the category of symmetric spectra such that the homology groups of a complex C become isomorphic to the homotopy groups of HC. Since H is lax monoidal, if A is a dg $\mathbb{Z}$-algebra, then HA is naturally an algebra in the category of symmetric spectra and if M is an A-module, then HM becomes an HA-module. More generally, if $\mathcal{A}$ is a dg category over $\mathbb{Z}$, then $H\mathcal{A}$ becomes a *spectral category*, *i.e.* a category enriched in symmetric spectra, *cf.* [138], [140]. Each $\mathcal{A}$-module M then gives rise to a *spectral module* HM over $H\mathcal{A}$. The spectral modules over a spectral category form a Quillen model category [140].

Recall that if $\mathcal{L}$ and $\mathcal{M}$ are Quillen model categories, a *Quillen adjunction* is given by a pair of adjoint functors $L\colon \mathcal{L} \to \mathcal{M}$ and $R\colon \mathcal{M} \to \mathcal{L}$ such that L preserves cofibrations and R fibrations. Such a pair induces an adjoint pair between the homotopy categories of $\mathcal{L}$ and $\mathcal{M}$. If the induced functors are equivalences, then (L, R) is a *Quillen equivalence*. The model categories $\mathcal{L}$ and $\mathcal{M}$ are *Quillen equivalent* if they are linked by a chain of Quillen equivalences.

It was shown by A. Robinson [129], *cf.* also [139], that for an ordinary ring R, the unbounded derived category of R-modules is equivalent to the homotopy category of spectral modules over HR. This result is generalized and refined as follows:

Theorem 3.13 (Shipley [140]). *If $\mathcal{A}$ is a dg category over $\mathbb{Z}$, the model categories of dg $\mathcal{A}$-modules and of spectral modules over $H\mathcal{A}$ are Quillen equivalent.*

This allows us to define two small dg categories $\mathcal{A}$ and $\mathcal{B}$ to be *topologically Morita equivalent* if their categories of spectral modules are Quillen equivalent.

Proposition 3.14 ([38]). *Let $\mathcal{A}$ and $\mathcal{B}$ be two dg rings. Then statement* a) *implies* b) *and* b) *implies* c):

a) *$\mathcal{A}$ and $\mathcal{B}$ are dg Morita equivalent.*

b) *$\mathcal{A}$ and $\mathcal{B}$ are topologically Morita equivalent.*

c) *$\mathcal{D}(\mathcal{A})$ is triangle equivalent to $\mathcal{D}(\mathcal{B})$.*

It is remarkable that in general, these implications are strict. Examples which show this were obtained in recent joint work by D. Dugger and B. Shipley [38], *cf.* also [141]. To show that c) does not imply b), they invoke Schlichting's example [135]: Let p be an odd prime. The module categories over $A' = \mathbb{Z}/p^2$ and $B' = (\mathbb{Z}/p)[\varepsilon]/\varepsilon^2$ are Frobenius categories. Their stable categories are triangle equivalent (both are equivalent to the category of $\mathbb{Z}/p$-vector spaces with the identical suspension and the split triangles) but the K-theories associated with the stable module categories are not isomorphic. Since K-theory is preserved under topological Morita equivalence (*cf.* Section 5.2 below), the dg algebras A and B associated (*cf.* Section 3.6) with the canonical generators (corresponding to the one-dimensional vector space over $\mathbb{Z}/p$) of the stable categories of A' and B' cannot be topologically Morita equivalent.

To show that b) does not imply a), Dugger and Shipley consider two dg algebras A and B with homology isomorphic to $\mathbb{Z}/2 \oplus \mathbb{Z}/2[2]$. The isomorphism classes of such algebras in the homotopy category of dg $\mathbb{Z}$-algebras are parametrized by the Hochschild cohomology group $HH^4_{\mathbb{Z}}(\mathbb{Z}/2, \mathbb{Z}/2)$. Their isomorphism classes in the homotopy category of $\mathbb{S}$-algebras are parametrized by the topological Hochschild cohomology group $THH^4_{\mathbb{S}}(\mathbb{Z}/2, \mathbb{Z}/2)$ as shown in [97]. The computation of the Hochschild cohomology group $HH^4_{\mathbb{Z}}(\mathbb{Z}/2, \mathbb{Z}/2)$ is elementary and, thanks to Franjou–Lannes–Schwartz' work [50], the topological Hochschild cohomology algebra

$$THH^*_{\mathbb{S}}(\mathbb{Z}/2, \mathbb{Z}/2)$$

is known. Dugger–Shipley then conclude by exhibiting a non-trivial element in the kernel of the canonical map

$$\Phi\colon HH^4_{\mathbb{Z}}(\mathbb{Z}/2, \mathbb{Z}/2) \to THH^4_{\mathbb{S}}(\mathbb{Z}/2, \mathbb{Z}/2).$$

The explicit description of the two algebras is given in [140], [36], [38]. The appearance of torsion in these examples is unavoidable: for dg algebras over the rationals, statements a) and b) above are equivalent [38].

4. The homotopy category of small dg categories

4.1. Introduction. Invariants like K-theory, Hochschild homology, cyclic homology… naturally extend from k-algebras to dg categories (*cf.* Section 5). In analogy with the case of ordinary k-algebras, these extended invariants are preserved under dg Morita equivalence. However, unlike the module category over a k-algebra, the derived category of a dg category, even with its triangulated structure, does not contain enough information to compute the invariant (*cf.* the examples in Section 3.9). Our aim in this section is to present a category obtained from that of small dg categories by 'inverting the dg Morita equivalences'. It could be called the 'homotopy category of enhanced (idempotent complete) triangulated categories' [20] or the 'Morita homotopy category of small dg categories' Hmo, as in [145]. Invariants like K-theory and cyclic homology factor through the Morita homotopy category.

The Morita homotopy category very much resembles the category of small, idempotent complete, triangulated categories. In particular, it admits 'dg quotients' [34], which correspond to Verdier localizations. Like these, they are characterized by a universal property. The great advantages of the Morita homotopy category over that of small triangulated categories are that moreover, it admits *all* (*homotopy*) *limits and colimits* (like any homotopy category of a Quillen model category) and is *monoidal and closed*.

The Morita homotopy category Hmo is a full subcategory of the localization Hqe of the category of small dg categories with respect to the quasi-equivalences. The first step is therefore to analyze the larger category Hqe. Its morphism spaces are revealed by Toën's theorem 4.3 below.

4.2. Inverting quasi-equivalences. Let k be a commutative ring and dgcat_k the category of small dg k-categories as in Section 2.2. An analogue of the following theorem for simplicial categories is proved in [15].

Theorem 4.1 ([146]). *The category* dgcat_k *admits a structure of cofibrantly generated model category whose weak equivalences are the quasi-equivalences and whose fibrations are the dg functors* $F\colon \mathcal{A} \to \mathcal{B}$ *which induce componentwise surjections* $\mathcal{A}(X,Y) \to \mathcal{B}(FX,FY)$ *for all* X, Y *in* $\mathcal{A}$ *and such that, for each isomorphism* $v\colon F(X) \to Z$ *of* $H^0(\mathcal{B})$*, there is an isomorphism* u *of* $H^0(\mathcal{A})$ *with* $F(u) = v$.

This shows in particular that the *localization* Hqe of dgcat_k with respect to the quasi-equivalences has small Hom-sets and that we can compute morphisms from $\mathcal{A}$ to $\mathcal{B}$ in the localization as morphisms modulo homotopy from a cofibrant replacement $\mathcal{A}_{\mathrm{cof}}$ of $\mathcal{A}$ to $\mathcal{B}$ (note that all small dg categories are fibrant). In general, the cofibrant replacement $\mathcal{A}_{\mathrm{cof}}$ is not easy to compute with but if $\mathcal{A}(X,Y)$ is cofibrant in $\mathcal{C}(k)$ and the unit morphisms $k \to \mathcal{A}(X,X)$ admit retractions in $\mathcal{C}(k)$ for all objects X, Y of $\mathcal{A}$, for example if k is a field, then for $\mathcal{A}_{\mathrm{cof}}$, we can take the category with the same objects as $\mathcal{A}$ and whose morphism spaces are given by the 'reduced cobar-bar construction', *cf. e.g.* [34], [84]. The homotopy relation is then the one of [80, 3.3].

However, the morphism sets in the localization are much better described as follows: Consider two dg categories $\mathcal{A}$ and $\mathcal{B}$. If necessary, we replace $\mathcal{A}$ by a quasi-equivalent dg category so as to achieve that $\mathcal{A}$ is *k-flat*, *i.e.* the functor $\mathcal{A}(X, Y)\otimes?$ preserves quasi-isomorphisms for all X, Y of $\mathcal{A}$ (for example, we could take a cofibrant replacement of $\mathcal{A}$). Let $\mathrm{rep}(\mathcal{A}, \mathcal{B})$ be the full subcategory of the derived category $\mathcal{D}(\mathcal{A}^{\mathrm{op}} \otimes \mathcal{B})$ of $\mathcal{A}$-$\mathcal{B}$-bimodules formed by the bimodules X such that the tensor functor

$$? \overset{L}{\otimes}_{\mathcal{A}} X \colon \mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$$

takes the representable $\mathcal{A}$-modules to objects which are isomorphic to representable $\mathcal{B}$-modules. In other words, we require that $X(?, A)$ is quasi-isomorphic to a representable $\mathcal{B}$-module for each object A of $\mathcal{A}$. We call such a bimodule a *quasi-functor* since it yields a genuine functor

$$H^0(\mathcal{A}) \to H^0(\mathcal{B}).$$

We think of $\mathrm{rep}(\mathcal{A}, \mathcal{B})$ as the 'category of representations up to homotopy of $\mathcal{A}$ in $\mathcal{B}$'.

Theorem 4.2 (Toën [155]). *The morphisms from $\mathcal{A}$ to $\mathcal{B}$ in the localization of* dgcat_k *with respect to the quasi-equivalences are in natural bijection with the isomorphism classes of* $\mathrm{rep}(\mathcal{A}, \mathcal{B})$.

The theorem has been in limbo for some time, *cf.* [78, 2.3], [80], [34]. It is due to B. Toën, as a corollary of a much more precise statement: Recall from [66, Ch. 5] that each model category $\mathcal{M}$ admits a mapping space bifunctor

$$\mathrm{Map} \colon \mathrm{Ho}(\mathcal{M})^{\mathrm{op}} \times \mathrm{Ho}(\mathcal{M}) \to \mathrm{Ho}(\mathrm{Sset})$$

such that we have, for example, the natural isomorphisms

$$\pi_0(\mathrm{Map}(X, Y)) = \mathrm{Ho}(\mathcal{M})(X, Y).$$

The spaces Map may also be viewed as the morphism spaces in the Dwyer–Kan localization [44], [42], [43] of $\mathcal{M}$ with respect to the class of weak equivalences, *cf.* [43], [65]. Now let $\mathcal{R}(\mathcal{A}, \mathcal{B})$ be the category with the same objects as $\mathrm{rep}(\mathcal{A}, \mathcal{B})$ and whose morphisms are the quasi-isomorphisms of dg bimodules. Thus, the category $\mathcal{R}(\mathcal{A}, \mathcal{B})$ is a non-full subcategory of the category of dg bimodules $\mathcal{C}(\mathcal{A}^{\mathrm{op}} \otimes \mathcal{B})$.

Theorem 4.3 (Toën [155]). *There is a canonical weak equivalence of simplicial sets between* $\mathrm{Map}(\mathcal{A}, \mathcal{B})$ *and the nerve of the category* $\mathcal{R}(\mathcal{A}, \mathcal{B})$.

The theorem allows one to compute the homotopy groups of the *classifying space* $|\mathrm{dgcat}|$ *of dg categories*, which is defined as the nerve of the category of quasi-equivalences between dg categories. Of course, the connected components of this space are in bijection with the isomorphism classes of Hqe. Now let $\mathcal{A}$ be a small dg category. Then the fundamental group of $|\mathrm{dgcat}|$ at $\mathcal{A}$ is the group of automorphisms

of $\mathcal{A}$ in Hqe (*cf.* [120]). For example, if $\mathcal{A}$ is the category of bounded complexes of projective B-modules over an ordinary k-algebra B, then this group is the derived Picard group of B as studied in [133], [82], [166]. For the higher homotopy groups, we have the

Corollary 4.4 ([155]). a) *The group* $\pi_2(|\mathrm{dgcat}|, \mathcal{A})$ *is the group of invertible elements of the dg center of* $\mathcal{A}$ (= *its zeroth Hochschild cohomology group*).

b) *For* $i \geq 2$, *the group* $\pi_i(|\mathrm{dgcat}|, \mathcal{A})$ *is the* $(2-i)$*-th Hochschild cohomology of* $\mathcal{A}$.

4.3. Closed monoidal structure. As we have observed in Section 2.2, the category dgcat_k admits a tensor product $\otimes$ and an internal Hom-functor $\mathcal{H}om$. If $\mathcal{A}$ is cofibrant, then the functor $\mathcal{A}\otimes ?$ preserves weak equivalences so that the localization Hqe inherits a tensor product $\overset{L}{\otimes}$. However, the tensor product of two cofibrant dg categories is not cofibrant in general (in analogy with the fact that the tensor product of two non-commutative free algebras is not non-commutative free in general). By the adjunction formula

$$\mathcal{H}om(\mathcal{A}, \mathcal{H}om(\mathcal{B}, \mathcal{C})) = \mathcal{H}om(\mathcal{A} \otimes \mathcal{B}, \mathcal{C}),$$

it follows that even if $\mathcal{A}$ is cofibrant, the functor $\mathcal{H}om(\mathcal{A}, ?)$ cannot preserve weak equivalences in general and thus will not induce an internal Hom-functor in Hqe. Nevertheless, we have the

Theorem 4.5 ([34], [155]). *The monoidal category* $(\mathrm{Hqe}, \overset{L}{\otimes})$ *admits an internal* Hom-*functor* $\mathcal{RH}om$. *For two dg categories* $\mathcal{A}$ *and* $\mathcal{B}$ *such that* $\mathcal{A}$ *is* k*-flat, the dg category* $\mathcal{RH}om(\mathcal{A}, \mathcal{B})$ *is isomorphic in* Hqe *to the dg category* $\mathrm{rep}_{\mathrm{dg}}(\mathcal{A}, \mathcal{B})$, i.e. *the full subcategory of the dg category of* $\mathcal{A}$-$\mathcal{B}$*-bimodules whose objects are those of* $\mathrm{rep}(\mathcal{A}, \mathcal{B})$ *and which are cofibrant as bimodules.*

Thus we have equivalences (we suppose $\mathcal{A}$ k-flat)

$$H^0(\mathcal{RH}om(\mathcal{A}, \mathcal{B})) = H^0(\mathrm{rep}_{\mathrm{dg}}(\mathcal{A}, \mathcal{B})) \xrightarrow{\sim} \mathrm{rep}(\mathcal{A}, \mathcal{B}).$$

In terms of the internal Hom-functor $\mathcal{H}om$ of dgcat_k, we have

$$H^0(\mathcal{RH}om(\mathcal{A}, \mathcal{B})) = H^0(\mathcal{H}om(\mathcal{A}, \mathcal{B}))[\Sigma^{-1}],$$

where Σ is the set of morphisms $\phi\colon F \to G$ such that $\phi(A)$ is invertible in $H^0(\mathcal{B})$ for all objects A of $\mathcal{A}$, *cf.* [78].

Yet another description can be given in terms of A_∞-functors: Let $\mathcal{A}$ be a dg category such that the morphism spaces $\mathcal{A}(A, A')$ are cofibrant in $\mathcal{C}(k)$ and the unit maps $k \to \mathcal{A}(A, A)$ admit retractions in $\mathcal{C}(k)$ for all objects A, A' of $\mathcal{A}$. Then the dg category $\mathcal{RH}om(\mathcal{A}, \mathcal{B})$ is quasi-equivalent to the A_∞-category of (strictly unital) A_∞-functors from $\mathcal{A}$ to $\mathcal{B}$, *cf.* [91], [98], [104], [84]. Since $\mathcal{B}$ is a dg category, this A_∞-category is in fact a dg category.

An important point of classical Morita theory is that for two rings B, C, there is an equivalence between the category of B-C-bimodules and the category of coproduct preserving functors from the category of B-modules to that of C-modules (note that here and in what follows, we need to consider 'large' categories and should introduce universes to make our statements rigorous…). Similarly, if $\mathcal{A}$ is a small k-flat dg category, we consider the large dg category $\mathcal{D}_{\mathrm{dg}}(\mathcal{A})$: it is the full dg subcategory of $\mathcal{C}_{\mathrm{dg}}(\mathcal{A})$ whose objects are all the cofibrant dg modules. Thus we have an equivalence of categories

$$\mathcal{D}(\mathcal{A}) = H^0(\mathcal{D}_{\mathrm{dg}}(\mathcal{A})).$$

This shows that if $\mathcal{B}$ is another dg category, then each quasi-functor X in

$$\mathrm{rep}(\mathcal{D}_{\mathrm{dg}}(\mathcal{A}), \mathcal{D}_{\mathrm{dg}}(\mathcal{B}))$$

gives rise to a functor $\mathcal{D}(\mathcal{A}) \to \mathcal{D}(\mathcal{B})$. We say that the quasifunctor X *preserves coproducts* if this functor preserves coproducts.

Theorem 4.6 ([155]). *There is a canonical isomorphism in* Hqe

$$\mathcal{D}_{\mathrm{dg}}(\mathcal{A}^{\mathrm{op}} \otimes \mathcal{B}) \xrightarrow{\sim} \mathcal{RH}om_{\mathrm{c}}(\mathcal{D}_{\mathrm{dg}}(\mathcal{A}), \mathcal{D}_{\mathrm{dg}}(\mathcal{B})),$$

where $\mathcal{RH}om_{\mathrm{c}}$ denotes the full subcategory of $\mathcal{RH}om$ formed by the coproduct preserving quasifunctors.

If we apply this theorem to $\mathcal{B} = \mathcal{A}$ and compare the endomorphism algebras of the identity functors on both sides, we see that the Hochschild cohomology (*cf.* Section 5.4 below) of the small dg category $\mathcal{A}$ coincides with the Hochschild cohomology of the large dg category $\mathcal{D}_{\mathrm{dg}}(\mathcal{A})$, which is quite surprising. An analogous result for Grothendieck abelian categories (in particular, module categories) is due to T. Lowen and M. Van den Bergh [102].

4.4. Dg localizations, dg quotients, dg-derived categories. Let $\mathcal{A}$ be a small dg category. Let S be a set of morphisms of $H^0(\mathcal{A})$. Let us say that a morphism $R\colon \mathcal{A} \to \mathcal{B}$ of Hqe *makes S invertible* if the induced functor

$$H^0(\mathcal{A}) \to H^0(\mathcal{B})$$

takes each $s \in S$ to an isomorphism.

Theorem 4.7 ([155]). *There is a morphism $Q\colon \mathcal{A} \to \mathcal{A}[S^{-1}]$ of* Hqe *such that Q makes S invertible and each morphism R of* Hqe *which makes S invertible uniquely factors through Q.*

We call $\mathcal{A}[S^{-1}]$ the *dg localization of $\mathcal{A}$ at S*. Note that it is unique up to unique isomorphism in Hqe. It is constructed in [155] as a homotopy pushout

$$\begin{array}{ccc} \coprod_{s\in S} I & \longrightarrow & \mathcal{A} \\ \downarrow & & \downarrow \\ \coprod_{s\in S} k & \longrightarrow & \mathcal{A}[S^{-1}], \end{array}$$

where I denotes the dg k-category freely generated by one arrow $f\colon 0 \to 1$ of degree 0 with $df = 0$ and left vertical arrow is induced by the morphisms $I \to k$ which sends f to 1. The universal property of $Q\colon \mathcal{A} \to \mathcal{A}[S^{-1}]$ admits refined forms, namely, Q induces an equivalence of categories

$$\mathrm{rep}(\mathcal{A}[S^{-1}], \mathcal{B}) \xrightarrow{\sim} \mathrm{rep}_S(\mathcal{A}, \mathcal{B}),$$

an isomorphism of Hqe

$$\mathrm{rep}_{\mathrm{dg}}(\mathcal{A}[S^{-1}], \mathcal{B}) \xrightarrow{\sim} \mathrm{rep}_{\mathrm{dg},S}(\mathcal{A}, \mathcal{B}),$$

and a weak equivalence of simplicial sets

$$\mathrm{Map}(\mathcal{A}[S^{-1}], \mathcal{B}) \xrightarrow{\sim} \mathrm{Map}_S(\mathcal{A}, \mathcal{B}).$$

Here rep_S and $\mathrm{rep}_{\mathrm{dg},S}$ denote the full subcategories of quasi-functors whose associated functors $H^0(\mathcal{A}) \to H^0(\mathcal{B})$ make S invertible and Map_S the union of the connected components containing these quasi-functors.

An important variant is the following: Let $\mathcal{N}$ be a set of objects of $\mathcal{A}$. Let us say that a morphism $Q\colon \mathcal{A} \to \mathcal{B}$ of Hqe *annihilates* $\mathcal{N}$ if the induced functor

$$H^0(\mathcal{A}) \to H^0(\mathcal{B})$$

takes all objects of $\mathcal{N}$ to zero objects (*i.e.* objects whose identity morphism vanishes in $H^0(\mathcal{B})$).

Theorem 4.8 ([80], [34]). *There is a morphism $Q\colon \mathcal{A} \to \mathcal{A}/\mathcal{N}$ of* Hqe *which annihilates $\mathcal{N}$ and is universal among the morphisms annihilating $\mathcal{N}$.*

We call $\mathcal{A}/\mathcal{N}$ the *dg quotient of $\mathcal{A}$ by $\mathcal{N}$*. If $\mathcal{A}$ is k-flat (*cf.* Section 4.2), then $\mathcal{A}/\mathcal{N}$ admits a beautiful simple construction [34]: One adjoins to $\mathcal{A}$ a contracting homotopy for each object of $\mathcal{N}$. The general case can be reduced to this one or treated using orthogonal subcategories [80]. The dg quotient has refined universal properties analogous to those of the dg localization. In particular, the morphism $\mathcal{A} \to \mathcal{A}/\mathcal{N}$ induces an equivalence [34]

$$\mathrm{rep}(\mathcal{A}/\mathcal{N}, \mathcal{B}) \to \mathrm{rep}_{\mathcal{N}}(\mathcal{A}, \mathcal{B}),$$

where $\mathrm{rep}_{\mathcal{N}}$ denotes the full subcategory of quasi-functors whose associated functors $H^0(\mathcal{A}) \to H^0(\mathcal{B})$ annihilate $\mathcal{N}$.

Dg quotients yield functorial dg versions of Verdier localizations [160]. For example, if $\mathcal{E}$ is a small abelian (or, more generally, exact) category, we can take for $\mathcal{A}$ the dg category of bounded complexes $\mathcal{C}^b_{\mathrm{dg}}(\mathcal{E})$ over $\mathcal{E}$ and for $\mathcal{N}$ the dg category of bounded acyclic complexes $\mathcal{A}c^b_{\mathrm{dg}}(\mathcal{E})$. Then we obtain the *dg-derived category*

$$\mathcal{D}^b_{\mathrm{dg}}(\mathcal{E}) = \mathcal{C}^b_{\mathrm{dg}}(\mathcal{E})/\mathcal{A}c^b_{\mathrm{dg}}(\mathcal{E})$$

so that we have

$$\mathcal{D}^b(\mathcal{E}) = H^0(\mathcal{D}^b_{\mathrm{dg}}(\mathcal{E})).$$

More generally, every localization pair [80] (= Frobenius pair [134]) gives rise to a dg category. After taking the necessary set-theoretic precautions, we also obtain a dg-derived category

$$\mathcal{D}_{\mathrm{dg}}(\mathcal{E}) = \mathcal{C}_{\mathrm{dg}}(\mathcal{E})/\mathcal{A}c_{\mathrm{dg}}(\mathcal{E})$$

which refines the *unbounded* derived category of a k-linear Grothendieck abelian category $\mathcal{E}$. For a quasi-compact quasi-separated scheme X, let us write $\mathcal{D}_{\mathrm{dg}}(X)$ for $\mathcal{D}_{\mathrm{dg}}(\mathcal{E})$, where $\mathcal{E}$ is the Grothendieck abelian category of quasi-coherent sheaves on X. The following theorem shows that dg functors between dg derived categories are much more closely related to geometry than triangle functors between derived categories, *cf.* [23], [114].

Theorem 4.9 ([155]). *Let X and Y be quasi-compact separated schemes over k such that X is flat over* Spec k. *Then we have a canonical isomorphism in* Hqe

$$\mathcal{D}_{\mathrm{dg}}(X \times_k Y) \xrightarrow{\sim} \mathcal{R}\mathcal{H}om_{\mathrm{c}}(\mathcal{D}_{\mathrm{dg}}(X), \mathcal{D}_{\mathrm{dg}}(Y)),$$

where $\mathcal{R}\mathcal{H}om_{\mathrm{c}}$ denotes the full subcategory of $\mathcal{R}\mathcal{H}om$ formed by the coproduct preserving quasi-functors. Moreover, if X and Y are smooth and projective over Spec k, *we have a canonical isomorphism in* Hqe

$$\mathrm{par}_{\mathrm{dg}}(X \times_k Y) \xrightarrow{\sim} \mathcal{R}\mathcal{H}om(\mathrm{par}_{\mathrm{dg}}(X), \mathrm{par}_{\mathrm{dg}}(Y))$$

where $\mathrm{par}_{\mathrm{dg}}$ *denotes the full dg subcategory of $\mathcal{D}_{\mathrm{dg}}$ whose objects are the perfect complexes.*

4.5. Pretriangulated dg categories. Let $\mathcal{A}$ be a small dg category. We say that $\mathcal{A}$ is *pretriangulated* or *exact* if the image of the Yoneda functor

$$Z^0(\mathcal{A}) \to \mathcal{C}(\mathcal{A}), \quad X \mapsto X^\wedge$$

is stable under shifts in both directions and extensions (in the sense of the exact structure of Section 3.4). Equivalently, for all objects X, Y of $\mathcal{A}$ and all integers n, the object $X^\wedge[n]$ is isomorphic to $X[n]^\wedge$ and the cone over a morphism $f^\wedge\colon X^\wedge \to Y^\wedge$ is isomorphic to $C(f)^\wedge$ for unique objects $X[n]$ and $C(f)$ of $Z^0(\mathcal{A})$. If $\mathcal{A}$ is exact, then $Z^0(\mathcal{A})$ becomes a Frobenius subcategory of $\mathcal{C}(\mathcal{A})$ and $H^0(\mathcal{A})$ a triangulated subcategory of $\mathcal{H}(\mathcal{A})$. If $\mathcal{B}$ is an exact dg category and $\mathcal{A}$ an arbitrary dg category, then $\mathcal{H}om(\mathcal{A}, \mathcal{B})$ is exact (whereas $\mathcal{A} \otimes \mathcal{B}$ is not, in general).

If $\mathcal{A}$ is an arbitrary small dg category, there is a universal dg functor

$$\mathcal{A} \to \mathrm{pretr}(\mathcal{A})$$

to a pretriangulated dg category $\mathrm{pretr}(\mathcal{A})$, *i.e.* a functor inducing an equivalence

$$\mathcal{H}om(\mathcal{A}, \mathcal{B}) \xrightarrow{\sim} \mathcal{H}om(\mathrm{pretr}(\mathcal{A}), \mathcal{B})$$

for each exact dg category $\mathcal{B}$. The dg category $\mathrm{pretr}(\mathcal{A})$ is the *pretriangulated hull of* $\mathcal{A}$ constructed explicitly in [21], *cf.* also [34], [145].

For any dg category $\mathcal{A}$, the category $H^0(\mathrm{pretr}(\mathcal{A}))$ is equivalent to the triangulated subcategory of $\mathcal{H}\mathcal{A}$ generated by the representable dg modules. The functor pretr preserves quasi-equivalences and induces a left adjoint to the inclusion of the full subcategory of exact dg categories into the homotopy category Hqe. If $\mathcal{B}$ is pretriangulated, then so is $\mathcal{RH}om(\mathcal{A}, \mathcal{B})$ for each small dg category $\mathcal{A}$ and we have

$$\mathcal{RH}om(\mathrm{pretr}(\mathcal{A}), \mathcal{B}) \xrightarrow{\sim} \mathcal{RH}om(\mathcal{A}, \mathcal{B}).$$

4.6. Morita fibrant dg categories, exact sequences. A dg functor $F\colon \mathcal{A} \to \mathcal{B}$ between small dg categories is a *Morita morphism* if it induces an equivalence $\mathcal{D}(\mathcal{B}) \to \mathcal{D}(\mathcal{A})$. Each quasi-equivalence is a Morita morphism (*cf.* Section 3.8) and so is the canonical morphism $\mathcal{A} \to \mathrm{pretr}(\mathcal{A})$ from $\mathcal{A}$ to its pretriangulated hull.

Theorem 4.10 ([145]). *The category* dgcat_k *admits a structure of cofibrantly generated model category whose weak equivalences are the Morita morphisms and whose cofibrations are the same as those of the canonical model structure on* dgcat_k (cf. *Theorem* 4.1).

A dg category $\mathcal{A}$ is *Morita fibrant* (or *triangulated* in the terminology of [156]) iff it is fibrant with respect to this model structure. This is the case iff the canonical functor $H^0(\mathcal{A}) \to \mathrm{per}(\mathcal{A})$ is an equivalence iff $\mathcal{A}$ is pretriangulated and $H^0(\mathcal{A})$ is idempotent complete (*cf.* Section 3.5). We write $\mathcal{A} \to \mathrm{per}_{\mathrm{dg}}(\mathcal{A})$ for a fibrant replacement of $\mathcal{A}$ and then have

$$\mathrm{per}(\mathcal{A}) = H^0(\mathrm{per}_{\mathrm{dg}}(\mathcal{A})).$$

We write Hmo *for the localization* of dgcat_k with respect to the Morita morphisms. Then the functor $\mathcal{A} \mapsto \mathrm{per}_{\mathrm{dg}}(\mathcal{A})$ yields a right adjoint of the quotient functor $\mathrm{Hqe} \to \mathrm{Hmo}$ and induces an equivalence from Hmo onto the subcategory of Morita fibrant dg categories in Hqe, *cf.* [145]. The category Hmo is pointed: The dg category with one object and one morphism is both initial and terminal. Moreover, Hmo admits all finite coproducts (they are induced by the disjoint unions) and these are isomorphic to products.

Let

$$\mathcal{A} \xrightarrow{I} \mathcal{B} \xrightarrow{P} \mathcal{C} \tag{5}$$

be a sequence of Hqe such that $PI = 0$ in Hmo.

Theorem 4.11. *The following are equivalent:*

i) *In* Hmo, *I is a kernel of P and P a cokernel of I.*

ii) *The morphism I induces an equivalence of* $\mathrm{per}(\mathcal{A})$ *onto a thick subcategory of* $\mathrm{per}(\mathcal{B})$ *and P induces an equivalence of the idempotent closure* [8] *of the Verdier quotient with* $\mathrm{per}(\mathcal{C})$.

iii) *The functor I induces an equivalence of $\mathcal{D}(\mathcal{A})$ with a thick subcategory of $\mathcal{D}(\mathcal{B})$ and P identifies the Verdier quotient with $\mathcal{D}(\mathcal{C})$.*

The theorem is proved in [80]. The equivalence of ii) and iii) is a consequence of Thomason–Trobaugh's localization theorem [152], [108], [112]. We say that (5) is an *exact sequence* of Hmo if the conditions of the theorem hold. For example, if X is a quasi-compact quasi-separated scheme, $U \subset X$ a quasi-compact open subscheme and $Z = X \setminus U$, then the sequence

$$\mathrm{par}_{\mathrm{dg}}(X \text{ on } Z) \longrightarrow \mathrm{par}_{\mathrm{dg}}(X) \longrightarrow \mathrm{par}_{\mathrm{dg}}(U)$$

is an exact sequence of Hmo by the results of [152, Sect. 5], where $\mathrm{par}_{\mathrm{dg}}(X)$ denotes the dg quotient of the category of perfect complexes (viewed as a full dg subcategory of the category of complexes of $\mathcal{O}_X$-modules) by its subcategory of acyclic perfect complexes and $\mathrm{par}_{\mathrm{dg}}(X \text{ on } Z)$ the full subcategory of perfect complexes supported on Z.

4.7. Dg categories of finite type. Let $\mathcal{M}$ be a cofibrantly generated model category and I a small category. Recall that the category of functors $\mathcal{M}^I$ is again a cofibrantly generated model category (with the componentwise weak equivalences). Thus, the diagonal functor $\mathrm{Ho}(\mathcal{M}) \to \mathrm{Ho}(\mathcal{M}^I)$ admits a left adjoint, the *homotopy colimit functor*, and a right adjoint, the *homotopy limit functor*. An object X of $\mathcal{M}$ is *homotopically finitely presented* if, for each filtered direct system Y_i, $i \in I$, of $\mathcal{M}$, the canonical morphism

$$\mathrm{hocolim}\, \mathrm{Map}(X, Y_i) \to \mathrm{Map}(X, \mathrm{hocolim}\, Y_i)$$

is a weak equivalence of simplicial sets. The category $\mathcal{M}$ is *homotopically locally finitely presented* if, in $\mathrm{Ho}(\mathcal{M})$, each object is the homotopy colimit of a filtered direct system (in $\mathcal{M}$) of homotopically finitely presented objects.

For example [64], the category of dg algebras is homotopically locally finitely presented and a dg algebra is homotopically finitely presented iff, in the homotopy category, it is a retract of a non-commutative free graded algebra $k\langle x_1, \ldots, x_n\rangle$ endowed with a differential such that dx_i belongs to $k\langle x_1, \ldots, x_{i-1}\rangle$ for each $1 \leq i \leq n$. A dg category is *of finite type* if it is dg Morita equivalent to a homotopically finitely presented dg algebra.

Theorem 4.12 ([156]). *The category of small dg categories endowed with the canonical model structure whose weak equivalences are the Morita morphisms is homotopically locally finitely presented and a dg category is homotopically finitely presented iff it is of finite type.*

A dg category $\mathcal{A}$ is called *smooth* if the bimodule $(X, Y) \mapsto \mathcal{A}(X, Y)$ is perfect in $\mathcal{D}(\mathcal{A}^{\mathrm{op}} \overset{L}{\otimes} \mathcal{A})$. This property is invariant under dg Morita equivalence. The explicit

description of the homotopically finitely presented dg algebras shows that a dg category of finite type is smooth. Conversely [156], a dg category $\mathcal{A}$ is of finite type if it is smooth and *proper*, *i.e.* dg Morita equivalent to a dg algebra whose underlying complex of k-modules is perfect.

4.8. Moduli of objects in dg categories. Let T be a small dg category. In [156], B. Toën and M. Vaquié introduce and study the D^--stack (in the sense of [157]) of objects in T. By definition, this D^--stack is the functor

$$\mathcal{M}_T : \text{Scalg} \to \text{Sset}$$

which sends a simplicial commutative k-algebra A to the simplicial set

$$\text{Map}(T^{\text{op}}, \text{per}_{\text{dg}}(NA)),$$

where NA is the commutative dg k-algebra obtained from A by the Dold–Kan equivalence. They show that if T is a dg category of finite type, then this D^--stack is locally geometric and locally of finite presentation. Moreover, if $E : T \to \text{per}_{\text{dg}}(k)$ is a k-point of $\mathcal{M}_T$, then the tangent complex of $\mathcal{M}_T$ at E is given by

$$\mathcal{T}_{\mathcal{M}_T, E} \xrightarrow{\sim} \mathcal{RH}om(E, E)[1].$$

In particular, if E is quasi-isomorphic to a representable $x^\wedge$, then we have

$$\mathcal{T}_{\mathcal{M}_T, E} \xrightarrow{\sim} T(x, x)[1].$$

It follows that the restriction of $\mathcal{M}_T$ to the category of commutative k-algebras is a locally geometric ∞-stack in the sense of C. Simpson [144]. Here are three consequences derived from these results in [156]:

1) If T is a dg category over a field k and is smooth, proper and Morita fibrant, then the sheaf associated with the presheaf

$$R \mapsto \text{Aut}_{\text{Hqe}_R}(T \otimes_k R),$$

on the category of commutative k-algebras is a group scheme locally of finite type over k (*cf.* [166] for the case where T is an algebra).

2) If X is a smooth proper scheme over a commutative ring k, then the ∞-stack of perfect complexes on X is locally geometric.

3) If A is a (non-commutative) k-algebra over a field k, then the ∞-stack of bounded complexes of finite-dimensional A-modules is locally geometric if either A is the path algebra of a finite quiver or a finite-dimensional algebra of finite global dimension.

4.9. Dg orbit categories. Let $\mathcal{A}$ be a dg category and $F\colon \mathcal{A} \to \mathcal{A}$ an automorphism of $\mathcal{A}$ in Hqe. Let us assume for simplicity that F is given by a dg functor $\mathcal{A} \to \mathcal{A}$. The *dg orbit category* $\mathcal{A}/F^{\mathbb{Z}}$ has the same objects as $\mathcal{A}$ and the morphisms defined by

$$(\mathcal{A}/F^{\mathbb{Z}})(X, Y) = \bigoplus_{d\in\mathbb{Z}} \operatorname{colim}_n \mathcal{A}(F^n X, F^{n+d} Y).$$

The projection functor $P\colon \mathcal{A} \to \mathcal{A}/F^{\mathbb{Z}}$ is endowed with a canonical morphism $\phi\colon PF \to P$ which becomes invertible in $H^0(\mathcal{A}/F^{\mathbb{Z}})$ and the pair (P, ϕ) is the solution of a universal problem, *cf.* [83]. The category $H^0(\mathcal{A})/F^{\mathbb{Z}}$ is defined analogously. It is isomorphic to $H^0(\mathcal{A}/F^{\mathbb{Z}})$ and can be thought of as the 'category of orbits' of the functor F acting in $H^0(\mathcal{A})$.

Let us now assume that k is a field. Let Q be a quiver (= oriented graph) whose underlying graph is a Dynkin graph of type A, D or E. Let $\operatorname{mod} kQ$ be the abelian category of finite-dimensional representations of Q over k (*cf. e.g.* [52], [4]). Let $\mathcal{A} = \mathcal{D}^b_{\mathrm{dg}}(\operatorname{mod} Q)$ and $F\colon \mathcal{A} \to \mathcal{A}$ an automorphism in Hqe. We say that F *acts properly* if no indecomposable object of $\mathcal{D}^b(\operatorname{mod} kQ)$ is isomorphic to its image under F. For example, if Σ is the *Serre functor* of $\mathcal{A}$, defined by the bimodule

$$(X, Y) \mapsto \mathcal{H}om_k(\mathcal{A}(Y, X), k),$$

then Σ acts properly and, more generally, if S is the suspension functor, then $S^{-d}\Sigma$ acts properly for each $d \in \mathbb{N}$ unless Q is reduced to a point.

Theorem 4.13 ([83]). *If F acts properly, the orbit category $\mathcal{D}^b_{\mathrm{dg}}(\operatorname{mod} kQ)/F^{\mathbb{Z}}$ is Morita fibrant and thus $\mathcal{D}^b(\operatorname{mod} kQ)/F^{\mathbb{Z}}$ is canonically triangulated.*

In the particular case where $F = S^{-d}\Sigma$, the triangulated category $H^0(\mathcal{A}/F^{\mathbb{Z}})$ is Calabi–Yau [91] of CY-dimension d (*cf.* [83]). For $d = 1$, the category $H^0(\mathcal{A}/F^{\mathbb{Z}})$ is equivalent to the category of finite-dimensional projective modules over the preprojective algebra (*cf.* [56], [31], [128]) associated with the Dynkin graph underlying Q. For $d = 2$, one obtains the *cluster category* associated with the Dynkin graph. This category was introduced in [27] for type A and in [6] in the general case. It serves in the representation-theoretic approach (*cf. e.g.* [6], [26], [55]) to the study of cluster algebras [47], [48], [13], [49]. It seems likely [2] that if k is algebraically closed, the theorem yields almost all Morita fibrant dg categories whose associated triangulated categories have finite-dimensional morphism spaces and only finitely many isoclasses of indecomposables. In particular, those among these categories which are Calabi–Yau of fixed CY-dimension $d \gg 0$ are expected to be parametrized by the simply laced Dynkin diagrams.

5. Invariants

5.1. Additive invariants. Let Hmo_0 be the category with the same objects as Hmo and where morphisms $\mathcal{A} \to \mathcal{B}$ are given by elements of the Grothendieck group of

the triangulated category rep($\mathcal{A}$, $\mathcal{B}$). The composition is induced from that of Hmo. The category Hmo_0 is additive and endowed with a canonical functor $\mathrm{Hmo} \to \mathrm{Hmo}_0$ (*cf.* [22] for a related construction). One can show [145] that a functor F defined on Hmo with values in an additive category factors through $\mathrm{Hmo} \to \mathrm{Hmo}_0$ iff for each exact dg category $\mathcal{A}$ endowed with full exact dg subcategories $\mathcal{B}$ and $\mathcal{C}$ which give rise to a semi-orthogonal decomposition $H^0(\mathcal{A}) = (H^0(\mathcal{B}), H^0(\mathcal{C}))$ in the sense of [21], the inclusions induce an isomorphism $F(\mathcal{B}) \oplus F(\mathcal{C}) \xrightarrow{\sim} F(\mathcal{A})$. We then say that *$F$ is an additive invariant*. The most basic additive invariant is given by $F\mathcal{A} = K_0(\mathrm{per}\,\mathcal{A}))$. In Hmo_0, it becomes a corepresentable functor: $K_0(\mathrm{per}(\mathcal{A})) = \mathrm{Hmo}_0(k, \mathcal{A})$. As we will see below, the K-theory spectrum and all variants of cyclic homology are additive invariants. This is of interest since non-isomorphic objects of Hmo can become isomorphic in Hmo_0. For example, if k is an algebraically closed field, each finite-dimensional algebra of finite global dimension becomes isomorphic to a product of copies of k in Hmo_0 (*cf.* [78]) but it is isomorphic to such a product in Hmo only if it is semi-simple.

5.2. *K*-theory. Let $\mathcal{A}$ be a small dg k-category. Its K-theory $K(\mathcal{A})$ is defined by applying Waldhausen's construction [161] to a suitable category with cofibrations and weak equivalences: here, the category is that of perfect $\mathcal{A}$-modules, the cofibrations are the morphisms $i : L \to M$ of $\mathcal{A}$-modules which admit retractions as morphisms of graded $\mathcal{A}$-modules (*i.e.* the inflations of Section 3.4) and the weak equivalences are the quasi-isomorphisms. This construction can be improved so as to yield a functor K from dgcat_k to the homotopy category of spectra. As in [152], from Waldhausen's results [161] one then obtains the following

Theorem 5.1. a) [36] *The map $\mathcal{A} \mapsto K(\mathcal{A})$ yields a well-defined functor on* Hmo.

b) *Applied to the bounded dg-derived category $\mathcal{D}^b_{\mathrm{dg}}(\mathcal{E})$ of an exact category $\mathcal{E}$, the K-theory defined above agrees with Quillen K-theory.*

c) *The functor $\mathcal{A} \mapsto K(\mathcal{A})$ is an additive invariant. Moreover, each short exact sequence $\mathcal{A} \to \mathcal{B} \to \mathcal{C}$ of* Hmo (cf. *Section* 4.6) *yields a long exact sequence*

$$\cdots \longrightarrow K_i(\mathcal{A}) \longrightarrow K_i(\mathcal{B}) \longrightarrow K_i(\mathcal{C}) \longrightarrow \cdots \longrightarrow K_0(\mathcal{B}) \longrightarrow K_0(\mathcal{A}).$$

Part a) can be improved on: In fact, D. Dugger and B. Shipley show in [36] that K-theory is even preserved under *topological* Morita equivalence. Part c) can be improved on by defining *negative K-groups* and showing that the exact sequence extends indefinitely to the right. We refer to [134] for the most recent results, which include the case of dg categories. By combining part a) with Rickard's theorem 3.12, one obtains the invariance of the K-theory of rings under triangle equivalences between their derived categories. By combining a) and b), one obtains the invariance of the K-theory of abelian categories under equivalences between their derived categories which come from isomorphisms of Hmo (or, more generally, from topological Morita equivalences). In fact, according to A. Neeman's results [110] the K-theory of an

abelian category is even determined by the underlying triangulated category of its derived category, *cf.* [113] for a survey of his work.

Of course, any invariant defined for small triangulated categories applied to the perfect derived category yields an invariant of small dg categories. For example, Balmer–Witt groups (*cf.* [7] for a survey), defined for dg categories $\mathcal{A}$ endowed with a suitable involution $\mathcal{A} \xrightarrow{\sim} \mathcal{A}^{\mathrm{op}}$ in Hmo, yield such invariants.

5.3. Hochschild and cyclic homology. Let $\mathcal{A}$ be a small k-flat k-category. Following [106] the *Hochschild chain complex* of $\mathcal{A}$ is the complex concentrated in homological degrees $p \geq 0$ whose pth component is the sum of the

$$\mathcal{A}(X_p, X_0) \otimes \mathcal{A}(X_p, X_{p-1}) \otimes \mathcal{A}(X_{p-1}, X_{p-2}) \otimes \cdots \otimes \mathcal{A}(X_0, X_1),$$

where $X_0, \ldots, X_p$ range through the objects of $\mathcal{A}$, endowed with the differential

$$d(f_p \otimes \ldots \otimes f_0) = f_{p-1} \otimes \cdots \otimes f_0 f_p + \sum_{i=1}^{p} (-1)^i f_p \otimes \cdots \otimes f_i f_{i-1} \otimes \cdots \otimes f_0.$$

Via the cyclic permutations

$$t_p(f_{p-1} \otimes \cdots \otimes f_0) = (-1)^p f_0 \otimes f_{p-1} \otimes \cdots \otimes f_1$$

this complex becomes a precyclic chain complex and thus gives rise [77, Sect. 2] to a *mixed complex* $C(\mathcal{A})$ in the sense of [72], *i.e.* a dg module over the dg algebra $\Lambda = k[B]/(B^2)$, where B is of degree -1 and $dB = 0$. As shown in [72], all variants of cyclic homology [100] only depend on $C(\mathcal{A})$ considered in $\mathcal{D}(\Lambda)$. For example, the cyclic homology of $\mathcal{A}$ is the homology of the complex $C(\mathcal{A}) \overset{L}{\otimes}_{\Lambda} k$.

If $\mathcal{A}$ is a k-flat differential graded category, its mixed complex is the sum-total complex of the bicomplex obtained as the natural re-interpretation of the above complex. If $\mathcal{A}$ is an arbitrary dg k-category, its Hochschild chain complex is defined as the one of a k-flat (*e.g.* a cofibrant) resolution of $\mathcal{A}$.

Theorem 5.2 ([79], [80]). a) *The map* $\mathcal{A} \mapsto C(\mathcal{A})$ *yields an additive functor* $\mathrm{Hmo}_0 \to \mathcal{D}(\Lambda)$. *Moreover, each exact sequence of* Hmo (cf. *Section* 4.6) *yields a canonical triangle of* $\mathcal{D}(\Lambda)$.

b) *If A is a k-algebra, there is a natural isomorphism* $C(A) \xrightarrow{\sim} C(\mathrm{per}_{\mathrm{dg}}(A))$.

c) *If X is a quasi-compact separated scheme, there is a natural isomorphism* $C(X) \xrightarrow{\sim} C(\mathrm{par}_{\mathrm{dg}}(X))$, *where $C(X)$ is the cyclic homology of X in the sense of* [99], [163] *and* $\mathrm{par}_{\mathrm{dg}}(X)$ *the dg category defined in Section* 4.6.

The second statement in a) may be viewed as an excision theorem analogous to [164]. We refer to the recent proof [28] of Weibel's conjecture [162] on the vanishing of negative K-theory for an application of the theorem. The brave new algebra description of *topological* Hochschild (co-)homology [142] would suggest that it is

also preserved under topological Morita equivalence but no reference seems to exist as yet.

The endomorphism algebra $\mathcal{RH}om_\Lambda(k,k)$ is quasi-isomorphic to $k[u]$, where u is of degree 2 and $d(u)=0$. It acts on $C(\mathcal{A}) \overset{L}{\otimes}_\Lambda k$ and this action is made visible in the isomorphism

$$C(\mathcal{A}) \overset{L}{\otimes}_\Lambda k = C(\mathcal{A}) \otimes k[u]$$

where u is of degree 2 and the differential on the right hand complex is given by

$$d(x \otimes f) = d(x) \otimes f + (-1)^{|x|} xB \otimes uf.$$

The following 'Hodge–de Rham conjecture' is true for the dg category of perfect complexes on a smooth projective variety or over a finite-dimensional algebra of finite global dimension. It is wide open in the general case.

Conjecture 5.3 ([33], [90]). *If $\mathcal{A}$ is a smooth proper dg category over a field k of characteristic 0, then the homology of $C(\mathcal{A}) \otimes k[u]/(u^n)$ is a flat $k[u]/(u^n)$-module for all $n \geq 1$.*

5.4. Hochschild cohomology. Let $\mathcal{A}$ be a small cofibrant dg category. Its cohomological Hochschild complex $C(\mathcal{A},\mathcal{A})$ is defined as the product-total complex of the bicomplex whose 0th column is

$$\prod \mathcal{A}(X_0, X_0),$$

where X_0 ranges over the objects of $\mathcal{A}$, and whose pth column, for $p \geq 1$, is

$$\prod \mathcal{H}om_k(\mathcal{A}(X_{p-1}, X_p) \otimes \mathcal{A}(X_{p-2}, X_{p-1}) \otimes \cdots \otimes \mathcal{A}(X_0, X_1), \mathcal{A}(X_0, X_p))$$

where $X_0, \ldots, X_p$ range over the objects of $\mathcal{A}$. The horizontal differential is given by the Hochschild differential. This complex carries rich additional structure: As shown in [58], it is a B_∞-algebra, *i.e.* its bar construction carries, in addition to its canonical differential and comultiplication, a natural *multiplication* which makes it into a dg bialgebra. The B_∞-structure contains in particular the cup product and the Gerstenhaber bracket, which both descend to the Hochschild cohomology

$$HH^*(\mathcal{A},\mathcal{A}) = H^*C(\mathcal{A},\mathcal{A}).$$

The Hochschild cohomology is naturally interpreted as the homology of the complex

$$\mathcal{H}om(\mathbf{1}_\mathcal{A}, \mathbf{1}_\mathcal{A})$$

computed in the dg category $\mathcal{RH}om(\mathcal{A},\mathcal{A})$, where $\mathbf{1}_\mathcal{A}$ denotes the identity functor of $\mathcal{A}$ (*i.e.* the bimodule $(X,Y) \mapsto \mathcal{A}(X,Y)$). Then the cup product corresponds to the composition (whereas the Gerstenhaber bracket has no obvious interpretation). Each

$c \in HH^n(\mathcal{A}, \mathcal{A})$ gives rise to morphisms $cM\colon M \to M[n]$ of $\mathcal{D}(\mathcal{A})$, functorial in $M \in \mathcal{D}(\mathcal{A})$. Another interpretation links the Hochschild cohomology of $\mathcal{A}$ to the derived Picard group and to the higher homotopy groups of the category of quasi-equivalences between dg categories, *cf.* Section 4.2.

A natural way of obtaining the B_∞-algebra structure on $C(\mathcal{A}, \mathcal{A})$ is to consider the A_∞-category of A_∞-functors from $\mathcal{A}$ to itself [91], [98], [104]. Here, the B_∞-algebra $C(\mathcal{A}, \mathcal{A})$ appears as the endomorphism algebra of the identity functor (*cf.* [84]).

Note that $C(\mathcal{A}, \mathcal{A})$ is not functorial with respect to dg functors. However, if $F\colon \mathcal{A} \to \mathcal{B}$ is a fully faithful dg functor, it clearly induces a restriction map

$$F^*\colon C(\mathcal{B}, \mathcal{B}) \to C(\mathcal{A}, \mathcal{A})$$

and this map is compatible with the B_∞-structure. This can be used to construct [81] a morphism

$$\phi_X\colon C(\mathcal{B}, \mathcal{B}) \to C(\mathcal{A}, \mathcal{A})$$

in the homotopy category of B_∞-algebras associated with each dg $\mathcal{A}$-$\mathcal{B}$-bimodule X such that the functor

$$? \overset{L}{\otimes}_{\mathcal{A}} X\colon \operatorname{per}(\mathcal{A}) \to \mathcal{D}\mathcal{B}$$

is fully faithful. If moreover the functor $X \overset{L}{\otimes}_{\mathcal{B}} ?\colon \operatorname{per}(\mathcal{B}^{\mathrm{op}}) \to \mathcal{D}(\mathcal{A}^{\mathrm{op}})$ is fully faithful, then ϕ_X is an isomorphism. In particular, the Hochschild complex becomes a functor

$$\mathrm{Hmo}_{\mathrm{ff}}^{\mathrm{op}} \to \mathrm{Ho}(B_\infty),$$

where $\mathrm{Ho}(B_\infty)$ is the homotopy category of B_∞-algebras and $\mathrm{Hmo}_{\mathrm{ff}}$ the (non-full) subcategory of Hmo whose morphisms are the quasi-functors $X \in \operatorname{rep}(\mathcal{A}, \mathcal{B})$ such that

$$? \overset{L}{\otimes}_{\mathcal{A}} X\colon \operatorname{per}(\mathcal{A}) \to \operatorname{per}(\mathcal{B})$$

is fully faithful. We refer to [102] for the closely related study of the Hochschild complex of an abelian category.

Let us suppose that k is a field of characteristic 0. Endowed with the Gerstenhaber bracket the Hochschild complex $C(\mathcal{A}, \mathcal{A})$ becomes a differential graded Lie algebra and this Lie algebra 'controls the deformations of the A_∞-category $\mathcal{A}$', *cf. e.g.* [93]. Here the A_∞-structures (m_n), $n \geq 0$, may have a non-trivial term m_0. Some (but not all) Hochschild cocycles also correspond to deformations of $\mathcal{A}$ as an object of Hmo. To be precise, let $k[\varepsilon]$ be the algebra of dual numbers and consider the reduction functor

$$R\colon \mathrm{Hmo}_{k[\varepsilon]} \to \mathrm{Hmo}_k, \quad \mathcal{B} \mapsto \mathcal{B} \overset{L}{\otimes}_{k[\varepsilon]} k.$$

A *first order Morita deformation* of $\mathcal{A}$ is a pair $(\mathcal{A}', \phi)$ formed by a dg $k[\varepsilon]$-category $\mathcal{A}'$ and an isomorphism $\phi\colon R\mathcal{A}' \to \mathcal{A}$ of Hmo_k. An equivalence between such deformations is given by an isomorphism $\psi\colon \mathcal{A}' \to \mathcal{A}''$ such that $\phi' R\psi = \phi$. Then one can show [54] that the equivalence classes of first order Morita deformations of $\mathcal{A}$ are in

natural bijection with the classes $c \in HH^2(\mathcal{A}, \mathcal{A})$ such that the induced morphism $cP\colon P \to P[2]$ is nilpotent in $H^*\mathcal{H}om(P, P)$ for each perfect $\mathcal{A}$-module P. If $\mathcal{A}$ is proper or, more generally, if $H^n\mathcal{A}(?, X)$ vanishes for $n \gg 0$ for all objects X of $\mathcal{A}$, then this condition holds for all Hochschild 2-cocycles c. On the other hand, if $\mathcal{A}$ is given by the dg algebra $k[u, u^{-1}]$, where u is of degree 2 and $du = 0$, then it does not hold for the cocycle $u \in HH^2(\mathcal{A}, \mathcal{A})$.

5.5. Fine structure of the Hochschild complexes. The Hochschild cochain complex of a dg category carries a natural homotopy action of the little squares operad. This is the positive answer to a question by P. Deligne [29] which has been obtained, for example, in [105], [92], [14].... Hochschild cohomology acts on Hochschild homology and this action comes from a homotopy action of the Hochschild cochain complex, viewed as a homotopy algebra over the little squares, on the Hochschild chain complex. This is the positive answer to a series of conjectures due to B. Tsygan [158] and Tamarkin–Tsygan [149]. It has recently been obtained by B. Tsygan and D. Tamarkin [159]. Together, the two Hochschild complexes endowed with these structures yield a *non-commutative calculus* [150] analogous to the differential calculus on a smooth manifold. The link with classical calculus on smooth commutative manifolds is established through M. Kontsevich's formality theorem [89], [147] for Hochschild cochains and in [143] (*cf.* also [32]) for Hochschild chains.

Clearly, these finer structures on the Hochschild complexes are linked to the category of dg categories and its simplicial enrichment given by the Dwyer–Kan localization as developed in [155]. At the end of the introduction to [155], the reader will find a more detailed discussion of these links, *cf.* also [87]. A precise relationship is announced in [148].

5.6. Derived Hall algebras. Let $\mathcal{A}$ be a *finitary* abelian category, *i.e.* such that the underlying sets of $\mathcal{A}(X, Y)$ and $\mathrm{Ext}^1(X, Y)$ are finite for all objects X, Y of $\mathcal{A}$. The *Ringel–Hall algebra* $\mathcal{H}(\mathcal{A})$ is the free abelian group on the isomorphism classes of $\mathcal{A}$ endowed with the multiplication whose structure constants are given by the Hall numbers f^Z_{XY}, which count the number of subobjects of Z isomorphic to X and such that Z/X is isomorphic to Y, *cf.* [30] for a survey. Thanks to Ringel's famous theorem [126], [127], for each simply laced Dynkin diagram Δ, the *positive part* of the Drinfeld–Jimbo quantum group $U_q(\Delta)$ (*cf. e.g.* [73], [103]) is obtained as the (generic, twisted) Ringel–Hall algebra of the abelian category of finite-dimensional representations of a quiver $\vec{\Delta}$ with underlying graph Δ. Since Ringel's discovery, it was first pointed out by Xiao [165], *cf.* also [70], that an extension of the construction of the Ringel–Hall algebra to the derived category of the representations of $\vec{\Delta}$ might yield the *whole* quantum group. However, if one tries to mimic the construction of $\mathcal{H}(\mathcal{A})$ for a triangulated category $\mathcal{T}$ by replacing short exact sequences by triangles one obtains a multiplication which fails to be associative, *cf.* [70], [68]. It is remarkable that nevertheless, as shown by Peng–Xiao [115], [116], [117], the commutator associated with this multiplication yields the correct Lie algebra.

A solution to the problem of constructing an associative multiplication from the triangles has recently been proposed by B. Toën in [153]. He obtains an explicit formula for the structure constants ϕ_{XY}^Z of an associative multiplication on the rational vector space generated by the isomorphism classes of any triangulated category $\mathcal{T}$ which appears as the perfect derived category $\mathrm{per}(T)$ of a proper dg category T over a finite field k. The resulting $\mathbb{Q}$-algebra is the *derived Hall algebra* $\mathcal{DH}(T)$ *of* T. The formula for the structure constants reads as follows:

$$\phi_{XY}^Z = \sum_f |\operatorname{Aut}(f/Z)|^{-1} \prod_{i>0} |\operatorname{Ext}^{-i}(X, Z)|^{(-1)^i} |\operatorname{Ext}^{-i}(X, X)|^{(-1)^{i+1}},$$

where f ranges over the set of orbits of the group $\mathrm{Aut}(X)$ in the set of morphisms $f\colon X \to Z$ whose cone is isomorphic to Y, and $\mathrm{Aut}(f/Z)$ denotes the stabilizer of f under the action of $\mathrm{Aut}(X)$. The proof of associativity is inspired by methods from the study of higher moduli spaces [157], [155], [156] and by the homotopy theoretic approach to K-theory [120]. From the formula, it is immediate that $\mathcal{DH}(T)$ is preserved under triangle equivalences $\mathrm{per}(T) \xrightarrow{\sim} \mathrm{per}(T')$. Another consequence is that if $\mathcal{A}$ is the heart of a non-degenerate t-structure [10] on $\mathrm{per}(T)$, then the Ringel–Hall algebra of $\mathcal{A}$ appears as a subalgebra of $\mathcal{DH}(T)$. The derived Hall algebra of the derived category of representations of $\vec{\Delta}$ over a finite field appears closely related to the constructions of [70]. Its precise relation to the quantum group $U_q(\Delta)$ remains to be investigated.

Notice that like the K_0-group, the derived Hall algebra only depends on the underlying triangulated category of $\mathrm{per}(T)$. One would expect that geometric versions of the derived Hall algebra, as defined in [154, 3.3] will depend on finer data.

References

[1] Adams, J. Frank, A variant of E. H. Brown's representability theorem. *Topology* **10** (1971), 185–198.

[2] Amiot, Claire, Sur les catégories triangulées. Ph. D. thesis in preparation.

[3] Angeleri-Hügel, Lidia, Happel, Dieter, and Krause, Henning (eds.), *Handbook of tilting theory*. To appear.

[4] Auslander, M., Reiten, I., and Smalø, S., *Representation theory of Artin algebras*. Cambridge Stud. Adv. Math. 36, Cambridge University Press, Cambridge 1995.

[5] Avramov, Luchezar, and Halperin, Stephen, Through the looking glass: a dictionary between rational homotopy theory and local algebra. In *Algebra, algebraic topology and their interactions* (Stockholm, 1983), Lecture Notes in Math. 1183, Springer-Verlag, Berlin 1986, 1–27.

[6] Buan, Aslak Bakke, Marsh, Robert J., Reineke, Markus, Reiten, Idun, and Todorov, Gordana, Tilting theory and cluster combinatorics. *Adv. Math.*, to appear.

[7] Balmer, Paul, An introduction to triangular Witt groups and a survey of applications. In *Algebraic and arithmetic theory of quadratic forms*, Contemp. Math. 344, Amer. Math. Soc., Providence, RI, 2004, 31–58.

[8] Balmer, Paul, and Schlichting, Marco, Idempotent completion of triangulated categories. *J. Algebra* (2) **236** (2001), 819–834.

[9] Beĭlinson, A. A., Coherent sheaves on $\mathbf{P}^n$ and problems in linear algebra. Funktsional. Anal. i Prilozhen. **12** (3) (1978), 68–69.

[10] Beilinson, Alexander A., Bernstein, Joseph, and Deligne, Pierre, Analyse et topologie sur les espaces singuliers. *Astérisque* **100**, Soc. Math. France, 1982.

[11] Beĭlinson, A. A., Ginsburg, V. A., and Schechtman, V. V., Koszul duality. *J. Geom. Phys.* **5** (3) (1988), 317–350.

[12] Beilinson, Alexander, Ginzburg, Victor, and Soergel, Wolfgang, Koszul duality patterns in representation theory. *J. Amer. Math. Soc.* **9** (2) (1996), 473–527.

[13] Berenstein, Arkady, Fomin, Sergey, and Zelevinsky, Andrei, Cluster algebras. III. Upper bounds and double Bruhat cells. *Duke Math. J.* **126** (1) (2005), 1–52.

[14] Berger, Clemens, and Fresse, Benoit, Combinatorial operad actions on cochains. *Math. Proc. Cambridge Philos. Soc.* **137** (2004), 135–174.

[15] Bergner, Julie, A model category structure on the category of simplicial categories. *Trans. Amer. Math. Soc.*, to appear; math.AT/0406507.

[16] Bernšteĭn, I. N., Gel'fand, I. M., and Gel'fand, S. I., Algebraic vector bundles on $\mathbf{P}^n$ and problems of linear algebra. *Funktsional. Anal. i Prilozhen.* **12** (3) (1978), 66–67.

[17] Block, Jonathan, Duality and equivalence of module categories in noncommutative geometry I. arXiv:math.QA/0509284.

[18] Bökstedt, Marcel, and Neeman, Amnon, Homotopy limits in triangulated categories. *Compositio Math.* **86** (1993), 209–234.

[19] Bondal, A., and van den Bergh, M., Generators and representability of functors in commutative and noncommutative geometry. *Moscow Math. J.* **3** (1) (2003), 1–36, 258.

[20] Bondal, A. I., and Kapranov, M. M., Representable functors, Serre functors, and reconstructions. *Izv. Akad. Nauk SSSR Ser. Mat.* **53** (6) (1989), 1183–1205, 1337.

[21] —, Enhanced triangulated categories. *Mat. Sb.* **181** (5) (1990), 669–683; English transl. *Math. USSR-Sb.* **70** (1) (1990), 93-–107.

[22] Bondal, Alexey I., Larsen, Michael, and Lunts, Valery A., Grothendieck ring of pretriangulated categories. *Internat. Math. Res. Notices* **29** (2004), 1461–1495.

[23] Bondal, A., and Orlov, D., Derived categories of coherent sheaves. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 47–56.

[24] Bousfield, A. K., and Friedlander, E. M., Homotopy theory of Γ-spaces, spectra, and bisimplicial sets. In *Geometric applications of homotopy theory* (Proc. Conf., Evanston, Ill., 1977), II, Lecture Notes in Math. 658, Springer-Verlag, Berlin 1978, 80–130.

[25] Brown, E. H., Cohomology theories. *Ann. of Math.* **75** (1962), 467–484.

[26] Caldero, Philippe, and Chapoton, Frédéric, Cluster algebras as Hall algebras of quiver representations. arXiv:math.RT/0410187.

[27] Caldero, Philippe, Chapoton, Frédéric, and Schiffler, Ralf, Quivers with relations arising from clusters (A_n case). arXiv:math.RT/0401316.

[28] Cortiñas, G., Haesemeyer, C., Schlichting, M., and Weibel, C., Cyclic homology, cdh-cohomology and negative K-theory. Preprint, 2005.

[29] Deligne, Pierre, Letter to Stasheff, Gerstenhaber, May, Schechtman, Drinfeld. May 17, 1993.

[30] Deng, Bangming, and Xiao, Jie, On Ringel-Hall algebras. In *Representations of finite dimensional algebras and related topics in Lie theory and geometry*, Fields Inst. Commun. 40, Amer. Math. Soc., Providence, RI, 2004, 319–348.

[31] Dlab, Vlastimil, and Ringel, Claus Michael, The preprojective algebra of a modulated graph. In *Representation theory, II* (Proc. Second Internat. Conf., Carleton Univ., Ottawa, Ont., 1979), Lecture Notes in Math. 832, Springer-Verlag, Berlin 1980, 216–231.

[32] Dolgushev, Vasiliy, Covariant and equivariant formality theorems. math.QA/0307212.

[33] Drinfeld, Vladimir, DG categories. Talks at the Geometric Langlands seminar, University of Chicago, Fall 2002, notes by David Ben-Zvi.

[34] —, DG quotients of DG categories. *J. Algebra* **272** (2) (2004), 643–691.

[35] Drozd, Ju. A., Tame and wild matrix problems. In *Representation theory, II* (Proc. Second Internat. Conf., Carleton Univ., Ottawa, Ont., 1979), Lecture Notes in Math. 832, Springer-Verlag, Berlin 1980, 242–258.

[36] Dugger, Daniel, and Shipley, Brooke, K-theory and derived equivalences. *Duke Math. J.* **124** (3) (2004), 587–617.

[37] Dugger, Daniel, and Shipley, Brooke, Enrichments of additive model categories. arXiv: math.AT/0602107.

[38] —, Topological equivalences for differential graded algebras. Preprint, 2006.

[39] —, Postnikov towers of ring spectra. In preparation.

[40] Dwyer, W. G., and Greenlees, J. P. C., Complete modules and torsion modules. *Amer. J. Math.* **124** (2002), 199–220.

[41] Dwyer, W. G., Greenlees, J. P. C., and Iyengar, S., Duality in algebra and topology. arXiv:math.AT/0510247.

[42] Dwyer, W. G., and Kan, D. M., Calculating simplicial localizations. *J. Pure Appl. Algebra* **18** (1980), 17–35.

[43] —, Function complexes in homotopical algebra. *Topology* **19** (1980), 427–440.

[44] —, Simplicial localizations of categories. *J. Pure Appl. Algebra* **17** (1980), 267–284.

[45] Dwyer, W. G., and Spaliński, J., Homotopy theories and model categories. In *Handbook of algebraic topology*, North-Holland, Amsterdam 1995, 73–126.

[46] Elmendorf, A. D., Kriz, I., Mandell, M. A., and May, J. P., Rings, modules, and algebras in stable homotopy theory, Math. Surveys Monogr. 47, Amer. Math. Soc., Providence, RI, 1997.

[47] Fomin, Sergey, and Zelevinsky, Andrei, Cluster algebras. I. Foundations. *J. Amer. Math. Soc.* **15** (2) (2002), 497–529 (electronic).

[48] —, Cluster algebras. II. Finite type classification. *Invent. Math.* **154** (1) (2003), 63–121.

[49] —, Cluster algebras: notes for the CDM-03 conference. In *Current developments in mathematics*, 2003, Int. Press, Somerville, MA, 2003, 1–34.

[50] Franjou, Vincent, Lannes, Jean, and Schwartz, Lionel, Autour de la cohomologie de Mac Lane des corps finis. *Invent. Math.* **115** (3) (1994), 513–538.

[51] Franke, Jens, On the Brown representability theorem for triangulated categories. *Topology* **40** (4) (2001), 667–680.

[52] Gabriel, P., and Roiter, A. V., *Representations of finite-dimensional algebras*. Encyclopaedia Math. Sci. 73, Springer-Verlag, Berlin 1992.

[53] Gabriel, P., and Zisman, M.,*Calculus of fractions and homotopy theory*. Ergeb. Math. Grenzgeb. 35, Springer-Verlag, New York 1967.

[54] Geiß, Christof, and Keller, Bernhard, Infinitesimal deformations of derived categories. Oberwolfach talk, February 2005.

[55] Geiß, Christof, Leclerc, Bernard, and Schröer, Jan, Semicanonical bases and preprojective algebras. *Ann. Sci. École Norm. Sup.* (4) **38** (2) (2005), 193–253.

[56] Gel'fand, I. M., and Ponomarev, V. A., Model algebras and representations of graphs. *Funktsional. Anal. i Prilozhen.* **13** (3) (1979), 1–12.

[57] Gelfand, Sergei I., and Manin, Yuri I., Methods of homological algebra. Translated from the 1988 Russian original, Springer-Verlag, Berlin, 1996.

[58] Getzler, Ezra, and Jones, J. D. S., Operads, homotopy algebra, and iterated integrals for double loop spaces. hep-th/9403055.

[59] Grothendieck, Alexandre, Les dérivateurs, Manuscript, 1990. Edited electronically by M. Künzer, J. Malgoire and G. Maltsiniotis.

[60] Happel, Dieter, On the derived category of a finite-dimensional algebra. *Comment. Math. Helv.* **62** (3) (1987) 339–389.

[61] —, *Triangulated categories in the representation theory of finite-dimensional algebras*. London Math. Soc. Lecture Note Ser. 119, Cambridge University Press, Cambridge 1988.

[62] Heller, Alex, Stable homotopy categories. *Bull. Amer. Math. Soc.* **74** (1968), 28–63.

[63] —, Homotopy theories. *Mem. Amer. Math. Soc.* **71** (1988), no. 383.

[64] Hinich, Vladimir, Homological algebra of homotopy algebras. *Comm. Algebra* **25** (10) (1997), 3291–3323.

[65] Hirschhorn, Philip S., *Model categories and their localizations*. Math. Surveys Monogr. 99, Amer. Math. Soc., Providence, RI, 2003.

[66] Hovey, Mark, Model categories. Math. Surveys Monogr. 63, Amer. Math. Soc., Providence, RI, 1999.

[67] Hovey, Mark, Shipley, Brooke, and Smith, Jeff, Symmetric spectra. *J. Amer. Math. Soc.* **13** (1) (2000), 149–208.

[68] Hubery, Andrew, From triangulated categories to Lie algebras: A theorem of Peng and Xiao. arXiv:math.RT/0502403.

[69] Illusie, Luc, Catégories dérivées et dualité: travaux de J.-L. Verdier. *Enseign. Math.* (2) **36** (3–4) (1990), 369–391.

[70] Kapranov, M., Heisenberg doubles and derived categories. *J. Algebra* **202** (2) (1998), 712–744.

[71] Kashiwara, Masaki, and Schapira, Pierre, *Categories and sheaves*. Grundlehren Math. Wiss. 332, Springer-Verlag, Berlin 2005.

[72] Kassel, Christian, Cyclic homology, comodules and mixed complexes. *J. Algebra* **107** (1987), 195–216.

[73] —, *Quantum groups*. Graduate Texts in Math. 155, Springer-Verlag, New York 1995.

[74] Keller, Bernhard, Chain complexes and stable categories. *Manuscripta Math.* **67** (4) (1990), 379–417.

[75] —, Derived categories and universal problems. *Comm. Algebra* **19** (1991), 699–747.

[76] —, Deriving DG categories. *Ann. Sci. École Norm. Sup.* (4) **27** (1) (1994), 63–102.

[77] —, Derived categories and their uses. In *Handbook of algebra*, Vol. 1, North-Holland, Amsterdam 1996, 671–701.

[78] —, Invariance and localization for cyclic homology of DG algebras. *J. Pure Appl. Algebra* **123** (1–3) (1998), 223–273.

[79] —, On the cyclic homology of ringed spaces and schemes. *Doc. Math.* **3** (1998), 231–259 (electronic).

[80] —, On the cyclic homology of exact categories. *J. Pure Appl. Algebra* **136** (1) (1999), 1–56.

[81] —, Derived invariance of higher structures on the Hochschild complex. Preprint, 2003.

[82] —, Hochschild cohomology and derived Picard groups. *J. Pure Appl. Algebra* **190** (1–3) (2004), 177–196.

[83] —, On triangulated orbit categories. *Doc. Math.* **10** (2005), 551–581.

[84] —, A-infinity algebras, modules and functor categories. arXiv:math.RT/ 0510508.

[85] Keller, Bernhard, and Vossieck, Dieter, Sous les catégories dérivées. *C. R. Acad. Sci. Paris Sér. I Math.* **305** (6) (1987), 225–228.

[86] Kelly, G. M., Chain maps inducing zero homology maps. *Proc. Cambridge Philos. Soc.* **61** (1965), 847–854.

[87] Kock, Joachim and Toën, Bertrand, Simplicial localization of monoidal structures, and a non-linear version of Deligne's conjecture. *Compositio Math.* **141** (2005), 253–261.

[88] König, Steffen, and Zimmermann, Alexander, *Derived equivalences for group rings*. With contributions by Bernhard Keller, Markus Linckelmann, Jeremy Rickard and Raphaël Rouquier, Lecture Notes in Math. 1685, Springer-Verlag, Berlin 1998.

[89] Kontsevich, Maxim, Deformation quantization of Poisson manifolds, I. Preprint of the IHÉS, October 1997, q-alg/9709040.

[90] —, Topological field theory for triangulated categories. Talk at the conference on K-Theory and Noncommutative Geometry, Institut Henri Poincaré, Paris, June 2004.

[91] —, *Triangulated categories and geometry*. Course at the École Normale Supérieure, Paris, Notes taken by J. Bellaïche, J.-F. Dat, I. Marin, G. Racinet and H. Randriambololona, 1998.

[92] Kontsevich, Maxim, and Soibelman, Yan, Deformations of algebras over operads and the Deligne conjecture. In *Conférence Moshé Flato 1999*, Vol. I, Math. Phys. Stud. 21, Kluwer Academic Publishers, Dordrecht 2000, 255–307.

[93] —, Homological mirror symmetry and torus fibrations. In *Symplectic geometry and mirror symmetry* (Seoul, 2000), World Scienfific Publishing, River Edge, NJ, 2001, 203–263.

[94] Krause, Henning, On Neeman's well generated triangulated categories. *Doc. Math.* **6** (2001), 121–126 (electronic).

[95] —, A Brown representability theorem via coherent functors. *Topology* **41** (4) (2002), 853–861.

[96] Krause, Henning, and Kussin, Dirk, Rouquier's theorem on representation dimension. arXiv:math.RT/0505055.

[97] Lazarev, A., Homotopy theory of A_∞ ring spectra and applications to MU-modules. *K-Theory* **24** (3) (2001), 243–281.

[98] Lefèvre-Hasegawa, Kenji, Sur les A_∞-catégories. Thèse de doctorat, Université Denis Diderot – Paris 7, November 2003, available at B. Keller's homepage; http://www.math.jussieu.fr/~keller/.

[99] Loday, Jean-Louis, Cyclic homology, a survey. In *Geometric and algebraic topology*, Banach Center Publ. 18, PWN, Warsaw 1986, 281–303.

[100] —, *Cyclic homology*. Second edition, Grundlehren Math. Wiss. 301, Springer-Verlag, Berlin 1998.

[101] Lowen, W., A generalization of the Gabriel-Popescu theorem. *J. Pure Appl. Algebra* **190** (1–3) (2004), 197–211.

[102] Lowen, Wendy Tor, and Van den Bergh, Michel, Hochschild cohomology of abelian categories and ringed spaces. arXive:math.KT/0405227.

[103] Lusztig, George, *Introduction to quantum groups*. Progr. Math. 110, Birkhäuser, Boston, MA, 1993.

[104] Lyubashenko, Volodymyr, Category of A_∞-categories. *Homology Homotopy Appl.* **5** (1) (2003), 1–48 (electronic).

[105] McClure, James E., and Smith, Jeffrey H., A solution of Deligne's Hochschild cohomology conjecture. In *Recent progress in homotopy theory* (Baltimore, MD, 2000), Contemp. Math. 293, Amer. Math. Soc., Providence, RI, 2002, 153–193.

[106] Mitchell, Barry, Rings with several objects. *Adv. in Math.* **8** (1972), 1–161.

[107] Neeman, Amnon, The derived category of an exact category. *J. Algebra* **135** (1990), 388–394.

[108] —, The connection between the K-theory localisation theorem of Thomason, Trobaugh and Yao, and the smashing subcategories of Bousfield and Ravenel. *Ann. Sci. École Norm. Sup.* (4) **25** (1992), 547–566.

[109] —, The Grothendieck duality theorem via Bousfield's techniques and Brown representability. *J. Amer. Math. Soc.* **9** (1996), 205–236.

[110] —, K-theory for triangulated categories I(A): homological functors. *Asian J. Math.* **1** (1997), 330–417.

[111] —, Brown representability for the dual. *Invent. Math.* **133** (1) (1998), 97–105.

[112] —, *Triangulated Categories*. Ann. of Math. Stud. 148, Princeton University Press, Princeton, NJ, 2001.

[113] —, The K-theory of triangulated categories. In *Handbook of K-theory* (ed. by E. Friedlander and D. Grayson), Vol. 2, Springer-Verlag, Berlin 2005, 1011–1078.

[114] Orlov, D. O., Equivalences of derived categories and $K3$ surfaces. *J. Math. Sci.* (*New York*) **84** (5) (1997), 1361–1381.

[115] Peng, Liangang and Xiao, Jie, A realization of affine Lie algebras of type $\widetilde{A}_{n-1}$ via the derived categories of cyclic quivers. In *Representation theory of algebras* (Cocoyoc, 1994), CMS Conf. Proc. 18, Amer. Math. Soc., Providence, RI, 1996, 539–554.

[116] —, Root categories and simple Lie algebras. *J. Algebra* **198** (1) (1997), 19–56.

[117] —, *Triangulated categories and Kac-Moody algebras*. *Invent. Math.* **140** (3) (2000), 563–603.

[118] Popesco, Nicolae, and Gabriel, Pierre, Caractérisation des catégories abéliennes avec générateurs et limites inductives exactes. *C. R. Acad. Sci. Paris Sér. I Math.* **258** (1964), 4188–4190.

[119] Porta, Marco, Ph. D. thesis in preparation.

[120] Quillen, Daniel, Higher algebraic K-theory. I. In *Algebraic K-theory*, I: Higher K-theories,Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972, Lecture Notes in Math. 341, Springer-Verlag, Berlin 1973, 85–147.

[121] —, *Homotopical algebra*. Lecture Notes in Math. 43, Springer-Verlag, Berlin 1967.

[122] Reiten, Idun, Tilting theory and quasitilted algebras. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 109–120.

[123] Rezk, Charles, Schwede, Stefan, and Shipley, Brooke, Simplicial structures on model categories and functors. *Amer. J. Math.* **123** (2001), 551–575.

[124] Rickard, Jeremy, Morita theory for derived categories. *J. London Math. Soc.* **39** (1989), 436–456.

[125] —, The abelian defect group conjecture. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 121–128.

[126] Ringel, Claus Michael, Hall algebras and quantum groups. *Invent. Math.* **101** (3) (1990), 583–591.

[127] —, Hall algebras revisited. In *Quantum deformations of algebras and their representations* (Ramat-Gan, 1991/1992; Rehovot, 1991/1992), Israel Math. Conf. Proc. 7, Bar-Ilan Univ., Ramat Gan 1993, 171–176.

[128] —, The preprojective algebra of a quiver. In *Algebras and modules, II* (Geiranger, 1996), CMS Conf. Proc. 24, Amer. Math. Soc., Providence, RI, 1998, 467–480.

[129] Robinson, Alan, The extraordinary derived category. *Math. Z.* **196** (2) (1987), 231–238.

[130] Roiter, A. V., Matrix problems. In *Proceedings of the International Congress of Mathematicians* (Helsinki, 1978), Vol. 1, Acad. Sci. Fennica, Helsinki 1980, 319–322.

[131] Rouquier, Raphaël, Dimensions of triangulated categories. arXiv:math.CT/0310134.

[132] —, Derived equivalences and finite dimensional algebras. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 191–221.

[133] Rouquier, Raphaël, and Zimmermann, Alexander, Picard groups for derived module categories. *Proc. London Math. Soc.* (3) **87** (1) (2003), 197–225.

[134] Schlichting, Marco, Negative K-theory of derived categories. *Math. Z.*, to appear.

[135] —, A note on K-theory and triangulated categories. *Invent. Math.* **150** (1) (2002), 111–116.

[136] Schwede, Stefan, Morita theory in abelian, derived and stable model categories. In *Structured ring spectra*, London Math. Soc. Lecture Note Ser. 315, Cambridge University Press, Cambridge 2004, 33–86.

[137] Schwede, Stefan, and Shipley, Brooke, Algebras and modules in monoidal model categories. *Proc. London Math. Soc.* (3) **80** (2) (2000), 491–511.

[138] —, Equivalences of monoidal model categories. *Algebr. Geom. Topol.* **3** (2003), 287–334 (electronic).

[139] —, Stable model categories are categories of modules. *Topology* **42** (1) (2003), 103–153.

[140] Shipley, Brooke, HZ-algebra spectra are differential graded algebras. arXiv:math.AT/0209215.

[141] —, Morita theory in stable homotopy theory. In *Handbook of Tilting Theory*, ed. by L. Angeleri-Hügel, D. Happel and H. Krause, to appear.

[142] —, Symmetric spectra and topological Hochschild homology. *K-Theory* **19** (2000), 155–183.

[143] Shoikhet, Boris, A proof of the Tsygan formality conjecture for chains. *Adv. Math.* **179** (1) (2003), 7–37.

[144] Simpson, Carlos, Algebraic (geometric) n-stacks. arXiv:math.AG/9609014.

[145] Tabuada, Gonçalo, Invariants additifs de dg-catégories. *Internat. Math. Res. Notices* **53** (2005), 3309–3339.

[146] —, Une structure de catégorie de modèles de Quillen sur la catégorie des dg-catégories. *C. R. Acad. Sci. Paris Sér. I Math.* **340** (1) (2005), 15–19.

[147] Tamarkin, D. E., Another proof of M. Kontsevich's formality theorem. Preprint, math.QA/9803025.

[148] —, What do DG categories form? Talk at the Geometric Landlands seminar, Chicago, October 2005, preprint in preparation.

[149] Tamarkin, D., and Tsygan, B., Noncommutative differential calculus, homotopy BV algebras and formality conjectures. *Methods Funct. Anal. Topology* **6** (2) (2000), 85–100.

[150] —, The ring of differential operators on forms in noncommutative calculus. In *Graphs and patterns in mathematics and theoretical physics*, Proc. Sympos. Pure Math. 73, Amer. Math. Soc., Providence, RI, 2005, 105–131.

[151] Alonso Tarrío, Leovigildo, Jeremías López, Ana, and Souto Salorio, María José, Localization in categories of complexes and unbounded resolutions. *Canad. J. Math.* **52** (2) (2000), 225–247.

[152] Thomason, Robert W., and Trobaugh, Thomas F., Higher algebraic K-theory of schemes and of derived categories, In *The Grothendieck Festschrift*, Vol. 3, Progr. Math. 88, Birkhäuser, Boston, MA, 1990, 247–435.

[153] Toën, Bertrand, Derived Hall algebras. arXiv:math.QA/0501343.

[154] —, Higher and derived stacks: a global overview. Preprint, available at the author's homepage.

[155] —, The homotopy theory of dg-categories and derived Morita theory. arXiv:math.AG/0408337.

[156] Toën, Bertrand, and Vaquié, M., Moduli of objects in dg-categories. arXiv:math.AG/0503269.

[157] Toën, Bertrand, and Vezzosi, Gabriele, Homotopical algebraic geometry II: Geometric stacks and applications. *Mem. Amer. Math. Soc*, to appear.

[158] Tsygan, B., Formality conjectures for chains. In *Differential topology, infinite-dimensional Lie algebras, and applications*, Amer. Math. Soc. Transl. Ser. (2) 194, Amer. Math. Soc., Providence, RI, 1999, 261–274.

[159] —, *Noncommutative differential calculus*. Course during the special period on K-Theory and Noncommutative Geometry, Institut Henri Poincaré, Paris, Spring 2004.

[160] Verdier, Jean-Louis, Des catégories dérivées des catégories abéliennes. *Astérisque* **239**, Société Mathématique de France, 1996.

[161] Waldhausen, Friedhelm, Algebraic K-theory of spaces. In *Algebraic and geometric topology* (New Brunswick, N.J., 1983), Springer-Verlag, Berlin 1985, 318–419.

[162] Weibel, Charles, K-theory and analytic isomorphisms. *Invent. Math.* **61** (1980), 177–197.

[163] —, Cyclic homology for schemes. *Proc. Amer. Math. Soc.* **124** (6) (1996), 1655–1662.

[164] Wodzicki, Mariusz, Excision in cyclic homology and in rational algebraic K-theory. *Ann. of Math.* (2) **129** (3) (1989), 591–639.

[165] Xiao, Jie, Hall algebra in a root category. SFB preprint 95-070, Bielefeld University, 1995.

[166] Yekutieli, Amnon, The derived Picard group is a locally algebraic group. *Algebr. Represent. Theory* **7** (2004), 53–57.

Université Denis Diderot – Paris 7, UFR de Mathématiques, Institut de Mathématiques, UMR 7586 du CNRS, Case 7012, 2 place Jussieu, 75251 Paris Cedex 05, France
E-mail: keller@math.jussieu.fr

Derived equivalences and finite dimensional algebras

Raphaël Rouquier

Abstract. We discuss the homological algebra of representation theory of finite dimensional algebras and finite groups. We present various methods for the construction and the study of equivalences of derived categories: local group theory, geometry and categorifications.

Mathematics Subject Classification (2000). 20C20, 18E30, 16E10.

Keywords. Derived categories, finite groups, representations, algebraic groups, categorification, homological dimension.

1. Introduction

This paper discusses derived equivalences, their construction and their use, for finite dimensional algebras, with a special focus on finite group algebras.

In a first part, we discuss Broué's abelian defect group conjecture and its ramifications. This is one of the deepest problem in the representation theory of finite groups. It is part of local representation theory, which aims to relate characteristic p representations of a finite group with representations of local subgroups (normalizers of non-trivial p-subgroups). We have taken a more functorial viewpoint in the definition of classical concepts (defect groups, subpairs,...).

In § 2.1.4, we present Alperin's conjecture, which gives a prediction for the number of simple representations, and Broué's conjecture, which is a much more precise prediction for the derived category, but does apply only to certain blocks (those with abelian defect groups).

We discuss in § 2.2 various types of equivalences that arise and present the crucial problem of lifting stable equivalences to derived equivalences.

In § 2.3, we present some local methods. We give a stronger version of the abelian defect group conjecture that can be approached inductively and reduced to the problem explained above of lifting stable equivalences to derived equivalences. Roughly speaking, in a minimal counterexample to that refinement of the abelian defect conjecture, there is a stable equivalence. Work of Rickard suggested to impose conditions on the terms of the complexes: they should be direct summands of permutation modules. We explain that one needs also to put conditions on the maps, that make the complexes look like complexes of chains of simplicial complexes.

There is no understanding on how to construct candidates complexes who would provide the derived equivalences expected by the abelian defect group conjecture in general. For finite groups of Lie type (in non-describing characteristic), we explain

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

(§ 2.4) Broué's idea that such complexes should arise as complexes of cohomology of Deligne–Lusztig varieties. We describe (§ 2.4.2) the Jordan decomposition of blocks (joint work with Bonnafé), as conjectured by Broué: Morita equivalences between blocks are constructed from the cohomology of Deligne–Lusztig varieties. For GL_n, every block is shown to be Morita equivalent to a unipotent block. This provides some counterpart to the Jordan decomposition of characters (Lusztig). In § 2.4.3 and 2.4.4, we explain the construction of complexes in the setting of the abelian defect conjecture. There are some delicate issues related to the choice of the Deligne–Lusztig variety and the extension of the action of the centralizer of a defect group to that of the normalizer. This brings braid groups and Hecke algebras of complex reflection groups.

In § 2.5, we explain how to view the problem of lifting stable equivalences to derived equivalences as a non-commutative version of the birational invariance of derived categories of Calabi–Yau varieties.

In § 2.6, we describe a class of derived equivalences which are filtered shifted Morita equivalences (joint work with Chuang). We believe these are the building bricks for most equivalences and the associated combinatorics should be interesting.

Part § 3 is devoted to some invariants of derived equivalences. In § 3.1, we explain a functorial approach to outer automorphism groups of finite dimensional algebras and deduce that their identity component is preserved under various equivalences. This functorial approach is similar to that of the Picard group of smooth projective schemes and we obtain also an invariance of the identity component of the product of the Picard group by the automorphism group, under derived equivalence.

In § 3.2, we explain how to transfer gradings through derived or stable equivalences. As a consequence, there should be very interesting gradings on blocks with abelian defect. This applies as well to Hecke algebras of type A in characteristic 0, where we obtain gradings which should be related to geometrical gradings.

Finally, in § 3.3, we explain the notion of dimension for triangulated categories, in particular for derived categories of algebras and schemes. This applies to answer a question of Auslander on the representation dimension and a question of Benson on Loewy length of group algebras.

Part § 4 is devoted to "categorifications". Such ideas have been advocated by I. Frenkel and have already shown their relevance in the work of Khovanov [57] on knot invariants. Our idea is that "classical" structures have natural higher counterparts. These act as symmetries of categories of representations or of sheaves.

In § 4.1, we explain the construction with Chuang of a categorification of $\mathfrak{sl}_2$ and we develop the associated "2-representation theory". There is an action on the sum of module categories of symmetric groups, and we deduce the existence of derived equivalences between blocks with isomorphic defect groups, using the general theory that provides a categorification of the adjoint action of the Weyl group. This applies as well to general linear groups, and gives a solution to the abelian defect group conjecture for symmetric and general linear groups.

In § 4.2, we define categorifications of braid groups. This is based on Soergel's bimodules.

I thank Cédric Bonnafé, Joe Chuang and Hyohe Miyachi for useful comments on a preliminary version of this paper.

2. Broué's abelian defect group conjecture

2.1. Introduction

2.1.1. Blocks. Let ℓ be a prime number. Let $\mathcal{O}$ be the ring of integers of a finite extension K of the field $\mathbb{Q}_\ell$ of ℓ-adic numbers and k its residue field.

Let G be a finite group. Modular representation theory is the study of the categories $\mathcal{O}G$-mod and kG-mod (finitely generated modules). The decomposition of Spec $Z(\mathcal{O}G)$ into connected components corresponds to the decomposition $Z(\mathcal{O}G) = \prod_b Z(\mathcal{O}G)b$, where b runs over the set of primitive idempotents of $Z(\mathcal{O}G)$ (the *block idempotents*). We have corresponding decompositions in *blocks* $\mathcal{O}G = \prod_b \mathcal{O}Gb$ and $\mathcal{O}G$-mod $= \bigoplus_b \mathcal{O}Gb$-mod.

Remark 2.1. One assumes usually that K is big enough so that KG is a product of matrix algebras over K (this will be the case if K contains the e-th roots of unity, where e is the exponent of G). Descent methods often allow a reduction to that case.

2.1.2. Defect groups. A *defect group* of a block $\mathcal{O}Gb$ is a minimal subgroup D of G such that $\mathrm{Res}^G_D = \mathcal{O}Gb \otimes_{\mathcal{O}Gb} -\colon D^b(\mathcal{O}Gb) \to D^b(\mathcal{O}D)$ is faithful (*i.e.*, injective on Hom's). Such a subgroup is an ℓ-subgroup and it is unique up to G-conjugacy.

The *principal block* $\mathcal{O}Gb_0$ is the one through which the trivial representation factors. Its defect groups are the Sylow ℓ-subgroups of G.

Defect groups measure the representation type of the block:

- kGb is simple if and only if $D = 1$.
- kGb-mod has finitely many indecomposable objects (up to isomorphism) if and only if the defect groups are cyclic.
- kGb is tame (*i.e.*, indecomposable modules are classifiable in a reasonable sense) if and only if the defect groups are cyclic or $\ell = 2$ and defect groups are dihedral, semi-dihedral or generalized quaternion groups.

2.1.3. Brauer correspondence. Let $\mathcal{O}Gb$ be a block and D a defect group. There is a unique block idempotent c of $\mathcal{O}N_G(D)$ such that the restriction functor $\mathrm{Res}^G_D = c\mathcal{O}Gb \otimes_{\mathcal{O}Gb} -\colon D^b(\mathcal{O}Gb) \to D^b(\mathcal{O}N_G(D)c)$ is faithful.

This correspondence provides a bijection between blocks of $\mathcal{O}G$ with defect group D and blocks of $\mathcal{O}N_G(D)$ with defect group D.

2.1.4. Conjectures. We have seen in § 2.1.3 that $D^b(\mathcal{O}G)$ embeds in $D^b(\mathcal{O}N_G(D)c)$. The *abelian defect conjecture* asserts that, when D is abelian, the categories are actually equivalent (via a different functor):

Conjecture 2.2 (Broué). If D is abelian, there is an equivalence $D^b(\mathcal{O}Gb) \xrightarrow{\sim} D^b(\mathcal{O}N_G(D)c)$.

A consequence of the conjecture is an isometry $K_0(KGb) \xrightarrow{\sim} K_0(KN_G(D)c)$ with good arithmetical properties (a *perfect isometry*). Note that the conjecture also carries homological information: if $\mathcal{O}Gb$ is the principal block and the equivalence sends the trivial module to the trivial module, we deduce that the cohomology rings of G and $N_G(D)$ are isomorphic, a classical and easy fact. It is unclear whether there should be some canonical equivalence in Conjecture 2.2.

Local representation theory is the study of the relation between modular representations and local structure of G. Alperin's conjecture asserts that the number of simple modules in a block can be computed in terms of local structure.

Conjecture 2.3 (Alperin). Assume $D \neq 1$. Then,

$$\text{rank } K_0(kGb) = \sum_{\mathcal{S}} (-1)^{l(\mathcal{S})+1} \text{ rank } K_0(kN_G(\mathcal{S})c_{\mathcal{S}})$$

where $\mathcal{S}$ runs over the conjugacy classes of chains of subgroups $1 < Q_1 < Q_2 < \cdots < Q_n \leq_G D, l(\mathcal{S}) = n \geq 1$ and $c_{\mathcal{S}}$ is the sum of the block idempotents of $N_G(\mathcal{S})$ corresponding to b.

Remark 2.4. We have stated here Knörr–Robinson's reformulation of the conjecture [58]. Note that the conjecture is expected to be compatible with ℓ-local properties of character degrees, equivariance, rationality (Dade, Robinson, Isaacs, Navarro). When D is abelian, Alperin's conjecture (and its refinements) follows immediately from Broué's conjecture. It would be extremely interesting to find a common refinement of Alperin and Broué's conjectures. For principal blocks, it should contain the description of the cohomology ring as stable elements in the cohomology ring of a Sylow subgroup.

2.2. Various equivalences. Let A and B be two symmetric algebras over a noetherian commutative ring $\mathcal{O}$.

2.2.1. Definitions. Let M be a bounded complex of finitely generated (A, B)-bimodules which are projective as A-modules and as right B-modules. Assume there is an (A, A)-bimodule R and a (B, B)-bimodule S with

$$M \otimes_B M^* \simeq A \oplus R \text{ as complexes of } (A, A)\text{-bimodules},$$
$$M^* \otimes_A M \simeq B \oplus S \text{ as complexes of } (B, B)\text{-bimodules}.$$

We say that M induces a

- *Morita equivalence* if M is concentrated in degree 0 and $R = S = 0$;
- *Rickard equivalence* if R and S are homotopy equivalent to 0 as complexes of bimodules;
- *derived equivalence* if R and S are acyclic;
- *stable equivalence* (of Morita type) if R and S are homotopy equivalent to bounded complexes of projective bimodules.

Note that Morita $\Rightarrow$ Rickard $\Rightarrow$ stable and Rickard $\Rightarrow$ derived. Note also that if there is a complex inducing a stable equivalence, then there is a bimodule inducing a stable equivalence. Finally, Rickard's theory says that if there is a complex inducing a derived equivalence, then there is a complex inducing a Rickard equivalence.

The definitions amount to requiring that $M \otimes_B -$ induces an equivalence

- (Morita) B-mod $\xrightarrow{\sim} A$-mod,
- (Rickard) $K^b(B\text{-mod}) \xrightarrow{\sim} K^b(A\text{-mod})$,
- (derived) $D^b(B) \xrightarrow{\sim} D^b(A)$,
- (stable) $B\text{-}\overline{\text{mod}} \xrightarrow{\sim} A\text{-}\overline{\text{mod}}$ (assuming $\mathcal{O}$ regular)

where $K^b(A\text{-mod})$ is the homotopy category of bounded complexes of objects of A-mod and $A\text{-}\overline{\text{mod}}$ is the stable category, additive quotient of A-mod by modules of the form $A \otimes_{\mathcal{O}} V$ with $V \in \mathcal{O}$-mod (it is equivalent to $D^b(A)/A$-perf when $\mathcal{O}$ is regular).

2.2.2. Stable equivalences. Stable equivalences arise fairly often in modular representation theory. For example, assume the Sylow ℓ-subgroups of G are TI, *i.e.*, given P a Sylow ℓ-subgroup, then $P \cap gPg^{-1} = 1$ for all $g \in G - N_G(P)$. Then, $M = \mathcal{O}G$ induces a stable equivalence between $\mathcal{O}G$ and $\mathcal{O}N_G(P)$, the corresponding functor is restriction (this is an immediate application of Mackey's formula). This restricts to a stable equivalence between principal blocks. Unfortunately, we do not know how to derive much numerical information from a stable equivalence.

A classical outstanding conjecture in representation theory of finite dimensional algebras is

Conjecture 2.5 (Alperin–Auslander). Assume $\mathcal{O}$ is an algebraically closed field. If A and B are stably equivalent, then they have the same number of isomorphism classes of simple non-projective modules.

A very strong generalization of Conjecture 2.5 is

Question 2.6. Let A and B be blocks with abelian defect groups and M a complex of (A, B)-bimodules inducing a stable equivalence. Assume K is big enough. Does there exist $\tilde{M}$ a complex of (A, B)-bimodules inducing a Rickard equivalence and such that M and $\tilde{M}$ are isomorphic in $(A \otimes B^{\text{opp}})\text{-}\overline{\text{mod}}$?

As will be explained in § 2.3.3, this is the key step for an inductive approach to Broué's conjecture.

Remark 2.7. There are examples of blocks with non abelian defect for which Question 2.6 has a negative answer, for example A the principal block of Suz(8), $\ell = 2$, and B the principal block of the normalizer of a Sylow 2-subgroup (TI case), cf. [17, §6]. A major problem with Question 2.6 and with Conjecture 2.2 is to understand the relevance of the assumption that the defect groups are abelian. Cf. § 3.2.2 for a possible idea.

2.3. Local theory. In an ideal situation, equivalences would arise from permutation modules or more generally, from chain complexes of simplicial complexes X acted on by the groups under consideration. Then, taking fixed points on X by an ℓ-subgroup Q would give rise to equivalences between blocks of the centralizers of Q. We would then have a compatible system of equivalences, corresponding to subgroups of the defect group. At the level of characters, Broué defined a corresponding notion of "isotypie" [17]: values of characters at ℓ-singular elements are related.

2.3.1. Subpairs. We explain here some classical facts.

A kG-module of the form $k\Omega$ where Ω is a G-set is a permutation module. An ℓ-permutation module is a direct summand of a permutation module and we denote by kG-lperm the corresponding full subcategory of kG-mod.

Suppose that Q is an ℓ-subgroup of G. We define the functor $\mathrm{Br}_Q\colon kG\text{-lperm} \to k(N_G(Q)/Q)\text{-lperm}$: $\mathrm{Br}_Q(M)$ is the image of M^Q in $M_Q = M/\sum_{x\in Q}(x-1)M$. If $M = k\Omega$, then $k(\Omega^Q) \xrightarrow{\sim} \mathrm{Br}_Q(M)$: the Brauer construction extends the fixed point construction on sets to ℓ-permutation modules. Note that this works only because Q is an ℓ-group and k has characteristic ℓ.

To deal with non principal blocks, we need to use Alperin–Broué's subpairs. A subpair of G is a pair (Q, e), where Q is an ℓ-subgroup of G and e a block idempotent of $kC_G(Q)$. If we restrict to the case where e is a principal block, we recover theory of ℓ-subgroups of G.

A maximal subpair is of the form (D, b_D), where D is a defect group of a block kGb and b_D is a block idempotent of $kC_G(D)$ such that $b_D c \neq 0$ (we say that (D, b_D) is a b-subpair). Fix such a maximal subpair. The $(kG, kN_G(D, b_D))$-bimodule $bkGb_D$ has, up to isomorphism, a unique indecomposable direct summand X with $\mathrm{Br}_{\Delta D}(X) \neq 0$. Here, we put $\Delta D = \{(x, x^{-1})\}_{x\in D} \leq D \times D^{\mathrm{opp}}$. More generally, given $\phi\colon Q \to R$, we put $\Delta_\phi(Q) = \{(x, \phi(x)^{-1})\}_{x\in Q} \leq Q \times R^{\mathrm{opp}}$.

We define the Brauer category $\mathcal{B}r(D, b_D)$: its objects are subpairs (Q, b_Q) with $Q \leq D$ and $b_Q \mathrm{Br}_{\Delta Q}(X) \neq 0$, and $\mathrm{Hom}((Q, b_Q), (R, b_R))$ is the set of $f \in \mathrm{Hom}(Q, R)$ such that there is $g \in G$ with $(Q^g, b_Q^g) \in \mathcal{B}r(D, b_D)$ and $f(x) = g^{-1}xg$ for all $x \in Q$.

Let $M \in kG$-lperm indecomposable. A vertex-subpair of M is a subpair (Q, b_Q) maximal such that $b_Q \mathrm{Br}_Q(M) \neq 0$ (such a subpair is unique up to conjugacy).

2.3.2. Splendid equivalences. Let G and H be two finite groups and b and b' two block idempotents of kG and kH.

The following Theorem [86], [92] shows that a stable equivalence corresponds to "local" Rickard equivalences, for complexes of ℓ-permutation modules.

Theorem 2.8. *Let M be an indecomposable complex of ℓ-permutation (kGb, kHb')-bimodules. Then M induces a stable equivalence between kGb and kHb' if and only if given (D, b_D) a maximal b-subpair, there is a maximal b'-subpair $(D', b'_{D'})$, an isomorphism $\phi\colon D \xrightarrow{\sim} D'$ inducing an isomorphism $\mathcal{B}r(D, b_D) \xrightarrow{\sim} \mathcal{B}r(D', b'_{D'})$ such that*

- *The indecomposable modules occurring in M have vertex-subpairs of the form $(\Delta_\phi(Q), b_Q \otimes b'_{\phi(Q)})$ for some $(Q, b_Q) \in \mathcal{B}r(D, b_D)$, with $(\phi(Q), b'_{\phi(Q)}) = \phi(Q, b_Q)$.*
- *For $1 \neq Q \leq D$, then $b_Q \cdot \mathrm{Br}_{\Delta_\phi Q}\, M \cdot b'_{\phi(Q)}$ induces a Rickard equivalence between $kC_G(Q)b_Q$ and $kC_H(Q)b'_{\phi(Q)}$, where $(Q, b_Q) \in \mathcal{B}r(D, b_D)$ and $(\phi(Q), b'_{\phi(Q)}) = \phi(Q, b_Q)$.*

Remark 2.9. In [83], Rickard introduced a notion of splendid equivalences for principal blocks (complexes of ℓ-permutation modules with diagonal vertices), later generalized by Harris [46] and Linckelmann [65]. Such equivalences were shown to induce equivalences for blocks of centralizers. In these approaches, an isomorphism between the defect groups of the two blocks involved was fixed a priori and vertex-subpairs were assumed to be "diagonal" with respect to the isomorphism. Theorem 2.8 shows it is actually easier and more natural to work with no a priori identification, and the property on vertex-subpairs is actually automatically satisfied.

The second part of the theorem (local Rickard equivalences $\Rightarrow$ stable equivalence) generalizes results of Alperin and Broué and is related to work of Bouc and Linckelmann.

Finally, a more general theory (terms need not be ℓ-permutation modules) has been constructed by Puig ("basic equivalences") [78].

Rickard proposed the following strengthening of Conjecture 2.2:

Conjecture 2.10. If D is abelian, there is a complex of ℓ-permutation modules inducing a Rickard equivalence between $\mathcal{O}Gb$ and $\mathcal{O}N_G(D)c$.

To the best of my knowledge, in all cases where Conjecture 2.2 is known to hold, then, Conjecture 2.10 is also known to hold.

Conjecture 2.10 is known to hold when D is cyclic [79], [62], [85]. In that case, one can construct a complex with length 2, but the longer complex originally constructed by Rickard might be more natural. The conjecture holds also when $D \simeq (\mathbb{Z}/2)^2$ [82], [63], [85]. In both cases, the representation type is tame. Note that there is no other ℓ-group P for which Conjecture 2.10 is known to hold for all $D \simeq P$.

Conjecture 2.10 holds when G is ℓ-solvable [35], [75], [47], when G is a symmetric group or a general linear group (cf. § 4.1; the describing characteristic case $G = \mathrm{SL}_2(\ell^n)$ is solved in [70]) and when G is a finite group of Lie type and $\ell \mid (q-1)$ (cf. § 2.4.3). There are many additional special groups for which the conjecture is known to hold (work of Gollan, Hida, Holloway, Koshitani, Kunugi, Linckelmann, Marcus, Miyachi, Okuyama, Rickard, Turner, Waki), cf. http://www.maths.bris.ac.uk/~majcr/adgc/adgc.html.

2.3.3. Gluing. Theorem 2.8 suggests an inductive approach to Conjecture 2.10: one should solve the conjecture for local subgroups (say, $C_G(Q)$, $1 \neq Q \leq D$) and glue the corresponding Rickard complexes. This would give rise to a complex inducing a stable equivalence, leaving us with the core problem of lifting a stable equivalence to a Rickard equivalence. Unfortunately, complexes are not rigid enough to allow gluing. This problem can be solved by using complexes endowed with some extra structure [86], [92]. The idea is to use complexes that have the properties of chain complexes of simplicial complexes: the key point is the existence of compatible splittings of the Brauer maps $M^Q \to M(Q)$. One can build an exact category of ℓ-permutation modules with compatible splittings of the Brauer maps. The subcategory of projective objects turns out to have a very simple description in terms of sets, and we use only this category. For simplicity, we restrict here to the case of principal blocks.

Let G be a finite group, ℓ a prime number, k an algebraically closed field of characteristic ℓ, b the principal block idempotent of kG, D a Sylow ℓ-subgroup of G and c the principal block idempotent of $H = N_G(D)$. We assume D is abelian. We denote by $Z_\ell(G)$ the Sylow ℓ-subgroup of $Z(G)$ and put $Z = \Delta Z_\ell(G)$.

Let G' be a finite group containing G as a normal subgroup, let $H' = N_{G'}(D)$ and $F = G'/G \xrightarrow{\sim} H'/H$. We assume F is an ℓ'-group, we put $N = \{(g,h) \in G' \times H'^{\mathrm{opp}} \mid (gG, hH^{\mathrm{opp}}) \in \Delta F\}$ and $\overline{N} = N/Z$.

Let $\mathcal{E}$ be the category of $\overline{N}$-sets whose point stabilizers are contained in $\Delta D/Z$. Let $\tilde{\mathcal{E}}$ be the Karoubian envelop of the linearization of $\mathcal{E}$ (objects are pairs (Ω, e) where Ω is a $\overline{N}$-set and e an idempotent of the monoid algebra of $\mathrm{End}_{\overline{N}}(\Omega)$). We have a faithful functor $\tilde{\mathcal{E}} \to k\overline{N}$-lperm, $(\Omega, e) \mapsto k(\Omega, e) := k\Omega e$.

We are now ready to state a further strengthening of Conjecture 2.2. For the inductive approach, it is important to take into account central ℓ-subgroups and ℓ'-automorphism groups.

Conjecture 2.11. There is a complex C of objects of $\tilde{\mathcal{E}}$ such that $\mathrm{Res}^{\overline{N}}_{G \times H^{\mathrm{opp}}} k(C)$ induces a Rickard equivalence between kGb and kHc.

We can also state a version of Question 2.6, for the pair (G', G):

Question 2.12. Let C be a complex of objects of $\tilde{\mathcal{E}}$ such that $\mathrm{Res}^{\overline{N}}_{G \times H^{\mathrm{opp}}} k(C)$ induces a stable equivalence between kGb and kHc. Is there a bounded complex

R of finitely generated projective $\bar{N}$-modules and a morphism $f\colon R \to k(C)$ such that $\mathrm{Res}^{\bar{N}}_{G\times H^{\mathrm{opp}}}\mathrm{cone}(f)$ induces a Rickard equivalence between kGb and kHc?

The following theorem reduces (a suitable version of) the abelian defect conjecture to (a suitable version of) the problem of lifting stable equivalences to Rickard equivalences.

Theorem 2.13. *Assume Question* 2.12 *has a positive answer for* $(N_{G'}(Q), C_G(Q))$ *for all non trivial subgroups* Q *of* D. *Then Conjecture* 2.11 *holds.*

The proof goes by building inductively (on the index of Q in D) a system of complexes for $N_{G'}(Q)$ and gluing them together. The key point is that, given a finite group Γ, the category of Γ-sets whose point stabilizers are non-trivial p-subgroups is locally determined. This allows us to manipulate objects of $\tilde{\mathcal{E}}$ as "sheaves".

2.4. Chevalley groups. We explain Broué's idea that complexes of cohomology of certain varieties should give rise to derived equivalences, for finite groups of Lie type.

2.4.1. Deligne–Lusztig varieties. Let $\boldsymbol{G}$ be a connected reductive algebraic group defined over a finite field and let F be an endomorphism of $\boldsymbol{G}$, a power F^d of which is a Frobenius endomorphism defining a structure over a finite field $\mathbb{F}_{q^d}$ for some $q \in \mathbb{R}_{>0}$. Let $G = \boldsymbol{G}^F$ be the associated finite group.

Let ℓ be a prime number with $\ell \nmid q$, K a finite extension of $\mathbb{Q}_l$, and $\mathcal{O}$ its ring of integers. We assume K is big enough.

Let $\boldsymbol{L}$ be an F-stable Levi subgroup of $\boldsymbol{G}$, $\boldsymbol{P}$ be a parabolic subgroup with Levi complement $\boldsymbol{L}$, and let $\boldsymbol{U}$ be the unipotent radical of $\boldsymbol{P}$. We define the Deligne–Lusztig variety

$$Y_{\boldsymbol{U}} = \{g\boldsymbol{U} \in \boldsymbol{G}/\boldsymbol{U} \mid g^{-1}F(g) \in \boldsymbol{U}\cdot F(\boldsymbol{U})\},$$

a smooth affine variety with a left action of $\boldsymbol{G}^F$ and a right action of $\boldsymbol{L}^F$ by multiplication. The corresponding complex of cohomology $R\Gamma_c(Y_{\boldsymbol{U}}, \mathcal{O})$ induces the Deligne–Lusztig induction functor $R^{\boldsymbol{G}}_{\boldsymbol{L}\subset\boldsymbol{P}}\colon D^b(\mathcal{O}\boldsymbol{L}^F) \to D^b(\mathcal{O}\boldsymbol{G}^F)$.

The effect of these functors on characters (*i.e.*, K_0's after extension to K) is a central tool for Deligne–Lusztig and Lusztig's construction of irreducible characters of G. It is important to also consider the finer invariant $\tilde{R}\Gamma_c(Y_{\boldsymbol{U}}, \mathcal{O})$, an object of $K^b(\mathcal{O}(\boldsymbol{G}^F \times (\boldsymbol{L}^F)^{\mathrm{opp}})\text{-lperm})$ which is quasi-isomorphic to $R\Gamma_c(Y_{\boldsymbol{U}}, \mathcal{O})$ [81], [87].

We put $X_{\boldsymbol{U}} = Y_{\boldsymbol{U}}/\boldsymbol{L}^F$ and denote by $\pi\colon Y_{\boldsymbol{U}} \to X_{\boldsymbol{U}}$ the quotient map.

Remark 2.14. One could use ordinary cohomology instead of the compact support version. One can conjecture that the two versions are interchanged by Alvis–Curtis duality: $(R\Gamma_c(Y_{\boldsymbol{U}}, \mathcal{O}) \otimes^{\boldsymbol{L}}_{\mathcal{O}\boldsymbol{L}^F} -) \circ D_{\boldsymbol{L}}$ and $D_{\boldsymbol{G}} \circ (R\Gamma(Y_{\boldsymbol{U}}, \mathcal{O})) \otimes^{\boldsymbol{L}}_{\mathcal{O}\boldsymbol{L}^F} -)$ should differ by a shift. This is known in the Harish-Chandra case, *i.e.*, when $\boldsymbol{P}$ is F-stable [24].

Let $\boldsymbol{T}_0 \subset \boldsymbol{B}_0$ be a pair consisting of an F-stable maximal torus and an F-stable Borel subgroup of $\boldsymbol{G}$. Let $\boldsymbol{U}_0$ be the unipotent radical of $\boldsymbol{B}_0$ and let $W = N_{\boldsymbol{G}}(\boldsymbol{T}_0)/\boldsymbol{T}_0$.

Let B^+ (resp. B) be the braid monoid (resp. group) of W. The canonical map $B^+ \to W$ has a unique section $w \mapsto \boldsymbol{w}$ that preserves lengths (it is not a group morphism!). We fix an F-equivariant morphism $\tau\colon B \to N_{\boldsymbol{G}}(\boldsymbol{T}_0)$ that lifts the canonical map $N_{\boldsymbol{G}}(\boldsymbol{T}_0) \to W$ [99]. Given $w \in W$, we put $\dot{w} = \tau(\boldsymbol{w})$. Let w_0 be the longest element of W and let $\pi = \boldsymbol{w}_0^2$, a central element of B.

Assume $\boldsymbol{L}$ above is a torus. We give a different model for $Y_{\boldsymbol{U}}$. Let $w \in W$ and $h \in \boldsymbol{G}$ such that $h^{-1}F(h) = \dot{w}$ and $\boldsymbol{U} = h\boldsymbol{U}_0h^{-1}$. Let

$$Y(w) = \{g\boldsymbol{U}_0 \in \boldsymbol{G}/\boldsymbol{U}_0 \mid g^{-1}F(g) \in \boldsymbol{U}_0\dot{w}\boldsymbol{U}_0\},$$

a variety with a left action of G and a right action of $\boldsymbol{T}_0^{wF}$ by multiplication. We have $\boldsymbol{L} = h\boldsymbol{T}_0h^{-1}$ and conjugation by h induces an isomorphism $\boldsymbol{L}^F \xrightarrow{\sim} \boldsymbol{T}_0^{wF}$. Right multiplication by h induces an isomorphism $Y_{\boldsymbol{U}} \xrightarrow{\sim} Y(w)$ compatible with the actions of G and $\boldsymbol{L}^F$. We have $\dim Y(w) = l(w)$. We write $Y_F(w)$ when the choice of F is important.

Given $w_1, \dots, w_r \in W$, we put

$$\begin{aligned} Y(w_1, \dots, w_r) = \{&(g_1\boldsymbol{U}_0, \dots, g_r\boldsymbol{U}_0) \in (\boldsymbol{G}/\boldsymbol{U}_0)^r \mid \\ &g_1^{-1}g_2 \in \boldsymbol{U}_0\dot{w}_1\boldsymbol{U}_0, \dots, g_{r-1}^{-1}g_r \in \boldsymbol{U}_0\dot{w}_{r-1}\boldsymbol{U}_0 \text{ and } g_r^{-1}F(g_1) \in \boldsymbol{U}_0\dot{w}_r\boldsymbol{U}_0\}. \end{aligned}$$

Up to a transitive system of canonical isomorphisms, $Y(w_1, \dots, w_r)$ depends only on the product $b = \boldsymbol{w}_1 \cdots \boldsymbol{w}_r \in B^+$ and we denote that variety by $Y(b)$ [36], [22].

2.4.2. Jordan decomposition. As a first step in his classification of (complex) irreducible characters of finite groups of Lie type, Lusztig established a *Jordan decomposition* of characters.

Let $(\boldsymbol{G}^*, F^*)$ be Langlands dual to $(\boldsymbol{G}, F)$. Then Lusztig defined a partition of the set $\mathrm{Irr}(G)$ of irreducible characters of G:

$$\mathrm{Irr}(G) = \coprod_{(s)} \mathrm{Irr}(G, (s))$$

where (s) runs over conjugacy classes of semi-simple elements of $(\boldsymbol{G}^*)^{F^*}$. The elements in $\mathrm{Irr}(G, 1)$ are the *unipotent characters*.

Furthermore, Lusztig constructed a bijection

$$\mathrm{Irr}(G, (s)) \xrightarrow{\sim} \mathrm{Irr}((C_{\boldsymbol{G}^*}(s)^*)^F, 1) \tag{1}$$

(assuming $C_{\boldsymbol{G}^*}(s)$ is connected). So, an irreducible character corresponds to a pair consisting of a semi-simple element in the dual and a unipotent character of the dual of the centralizer of that semi-simple element.

Broué and Michel [21] showed that the union of series corresponding to classes with a fixed ℓ'-part is a union of blocks: let t be a ℓ'-element of $(\boldsymbol{G}^*)^{F^*}$ and let

$$\mathrm{Irr}(G, (t))_\ell = \coprod_{(s)} \mathrm{Irr}(G, (s))$$

where (s) runs over conjugacy classes of semi-simple elements of $(\boldsymbol{G}^*)^{F^*}$ whose ℓ'-part is conjugate to t. Then $\mathrm{Irr}(G, (t))_\ell$ is a union of ℓ-blocks, and we denote by $B(G^F, (t))$ the corresponding factor of $\mathcal{O}G^F$.

Broué [18] conjectured that the decomposition (1) arises from a Morita equivalence (cf. also [48]). More, precisely, we have the following theorem [11, Theorem B'] obtained in joint work with C. Bonnafé (cf. also [23] for a detailed exposition). This was conjectured by Broué who gave a proof when t is regular [18].

Theorem 2.15 (Jordan decomposition of blocks). *Assume $C_{\boldsymbol{G}^*}(t)$ is contained in an F^*-stable Levi subgroup $\boldsymbol{L}^*$ of $\boldsymbol{G}^*$ with dual $\boldsymbol{L} \leq \boldsymbol{G}$. Let $\boldsymbol{P}$ be a parabolic subgroup of $\boldsymbol{G}$ with Levi complement $\boldsymbol{L}$ and unipotent radical $\boldsymbol{U}$. Let $d = \dim X_{\boldsymbol{U}}$ and let $\mathcal{F}_t = \pi_* \mathcal{O} \otimes_{\mathcal{O}\boldsymbol{L}^F} B(\boldsymbol{L}^F, (t))$.*

Then $H_c^i(X_{\boldsymbol{U}}, \mathcal{F}_t) = 0$ for $i \neq d$ and $H_c^d(X_{\boldsymbol{U}}, \mathcal{F}_t)$ induces a Morita equivalence between $B(G, (t))$ and $B(\boldsymbol{L}^F, (t))$.

The theorem reduces the study of blocks of finite groups of Lie type to the case of those associated to a quasi-isolated element t. When $\boldsymbol{L}^* = C_{\boldsymbol{G}^*}(t)$ is a Levi subgroup of $\boldsymbol{G}^*$, then $B(\boldsymbol{L}^F, (t))$ is isomorphic to $B(\boldsymbol{L}^F, 1)$.

As shown by Broué, the key point is the statement about the vanishing of cohomology. When $\boldsymbol{L}$ is a torus, this is [37, Theorem 9.8]. For the general case, two difficulties arise: there are no known good smooth compactifications of the varieties $X_{\boldsymbol{U}}$ and the locally constant sheaf $\mathcal{F}_t$ has wild ramification. We solve these issues as follows. Let $\overline{X}$ be the closure of $X_{\boldsymbol{U}}$ in $\boldsymbol{G}/\boldsymbol{P}$. We construct new varieties of Deligne–Lusztig type and commutative diagrams

$$\begin{array}{ccc} X_i & \xrightarrow{j_i} & Y_i \\ {\scriptstyle f_i'}\downarrow & & \downarrow{\scriptstyle f_i} \\ X_{\boldsymbol{U}} & \xrightarrow[j]{} & \overline{X} \end{array}$$

where Y_i is smooth, $Y_i - X_i$ is a divisor with normal crossings, and f_i is proper. We also construct tamely ramified sheaves $\mathcal{F}_i$ on X_i with the following properties:

- $\mathcal{F}_t$ is in the thick subcategory of the derived category of constructible sheaves on $\boldsymbol{X}_{\boldsymbol{U}}$ generated by the $Rf_{i*}'\mathcal{F}_i$
- $(Rj_{i*}\mathcal{F}_i)_{|f_i^{-1}(\overline{X} - X_{\boldsymbol{U}})} = 0$.

The first property follows from the following generation result of the derived category of a finite group of Lie type [11, Theorem A]:

Theorem 2.16. *The category of perfect complexes for $B(G, (t))$ is generated, as a thick subcategory, by the $R_{\boldsymbol{T}\subset\boldsymbol{B}}^{\boldsymbol{G}} B(\boldsymbol{T}^F, (t))$, where $\boldsymbol{T}$ runs over the F-stable maximal tori of $\boldsymbol{G}$ such that $t \in \boldsymbol{T}^*$ and $\boldsymbol{B}$ runs over the Borel subgroups of $\boldsymbol{G}$ containing $\boldsymbol{T}$.*

Remark 2.17. Note that the corresponding result for derived categories is true, under additional assumptions on $\boldsymbol{G}$ [13]: this is related to Quillen's Theorem, we need every elementary abelian ℓ-subgroup of G to be contained in an F-stable torus of $\boldsymbol{G}$.

Remark 2.18. Note that the Morita equivalence of Theorem 2.15 is not splendid in general. This issue is analyzed in [13].

Example 2.19. Let $\boldsymbol{G} = \mathrm{GL}_n(\overline{\mathbb{F}}_q)$ and $F\colon (x_{ij})_{1\le i,j\le n} \mapsto (x_{ij}^q)_{i,j}$. We have $G = \mathrm{GL}_n(\mathbb{F}_q)$, $\boldsymbol{G} = \boldsymbol{G}^*$ and $F^* = F$. Centralizers of semi-simple elements are Levi subgroups, so Theorem 2.15 gives a Morita equivalence between any block of a general linear group over $\mathcal{O}$ and a unipotent block.

2.4.3. Abelian defect conjecture. Let b be a block idempotent of $\mathcal{O}G$. Let (D, b_D) be a maximal b-subpair, let $H = N_G(D, b_D)$ and let $\boldsymbol{L} = C_{\boldsymbol{G}}(D)$. We assume D is abelian and $\boldsymbol{L}$ is a Levi subgroup of $\boldsymbol{G}$ (these are satisfied if $\ell \nmid |W|$).

Broué conjectured that the sought-for complex in Conjecture 2.10 should arise from Deligne–Lusztig varieties ([17, p. 81], [20, §1], [19, §VI]):

Conjecture 2.20 (Broué). There is a parabolic subgroup $\boldsymbol{P}$ of $\boldsymbol{G}$ with Levi complement $\boldsymbol{L}$ and unipotent radical $\boldsymbol{U}$, and a complex C inducing a Rickard equivalence between $\mathcal{O}Gb$ and $\mathcal{O}Hb_D$ such that $\mathrm{Res}_{G\times(\boldsymbol{L}^F)^{\mathrm{opp}}} C$ is isomorphic to $\tilde{R}\Gamma_c(Y_{\boldsymbol{U}}, \mathcal{O})b_D$.

This conjecture 2.20 is known to hold [76] when there is a choice of an F-stable parabolic subgroup $\boldsymbol{P}$ (case $\ell \mid (q-1)$). Then $Y_{\boldsymbol{U}}$ is 0-dimensional and the Deligne–Lusztig induction is the Harish-Chandra induction. The key steps in the proof are:

- Produce an action of the reflection group $H/\boldsymbol{L}^F$ from a natural action of the associated Hecke algebra. One needs to show that certain obstructions vanish.
- Identify a 2-cocycle of $H/\boldsymbol{L}^F$ with values in $\mathcal{O}^\times$.
- Compute the dimension of the KG-endomorphism ring.

2.4.4. Regular elements. As a first step, one should make Conjecture 2.20 more precise by specifying $\boldsymbol{P}$ and by defining the extension of the action of $C_G(D)$ to an action of H on $\tilde{R}\Gamma_c(Y_U, \mathcal{O})b_D$. These issues are partly solved and I will explain the best understood case where $\boldsymbol{L} = \boldsymbol{T}$ is a torus and $\mathcal{O}Gb$ is the principal block (cf. [22]). Assume as well $\ell \nmid (q-1)$. To simplify, assume further that F acts trivially on W ("split" case).

Note that $\boldsymbol{T}$ defines a conjugacy class $\mathcal{C}$ of W and the choice of $\boldsymbol{P}$ amounts to the choice of $w \in \mathcal{C}$ (defined from $\boldsymbol{P}$ as in § 2.4.1). Since $\boldsymbol{T} = C_{\boldsymbol{G}}(D)$, it follows that elements in $\mathcal{C}$ are Springer-regular. There is $w_d \in \mathcal{C}$ such that $(\boldsymbol{w}_d)^d = \pi$, where $d > 1$ is the order of w_d (a "good" regular element).

Given $w \in W$, we have a purely inseparable morphism

$$Y(w, w^{-1}w_0, w_0ww_0, w_0w^{-1}) \to Y(w^{-1}w_0, w_0ww_0, w_0w^{-1}, w)$$
$$(x_1, x_2, x_3, x_4) \mapsto (x_2, x_3, x_4, F(x_1)).$$

Via the canonical isomorphisms, this induces an endomorphism of $Y(\pi)$. This extends to an action of B^+ on $Y(\pi)$.

There is an embedding of $Y_F(w_d)$ as a closed subvariety of $Y_{F^d}(w_d, \dots, w_d)$ (d terms) given by

$$x \mapsto (x, F(x), \dots, F^{d-1}(x)).$$

The action of $C_{B^+}(\boldsymbol{w}_d)$ on $Y_{F^d}(\pi)$ restricts to an action on $Y_F(w_d)$. It induces an action of $C_B(\boldsymbol{w}^d)$ on $\tilde{R}\Gamma_c(Y(\boldsymbol{w}_d), \mathcal{O})$.

The group $H/C_G(D) \simeq C_W(w_d)$ is a complex reflection group and we denote by B_d its braid group. There is a morphism $B_d \to C_B(\boldsymbol{w}_d)$, uniquely defined up to conjugation by an element of the pure braid group of $C_W(w_d)$ (it is expected to be an isomorphism, and known to be such in a number of cases [8]).

Now, the conjecture is that, up to homotopy, the action of $\mathcal{O}(\boldsymbol{T}_0^{w_d F} \rtimes B_d)$ on $\tilde{R}\Gamma_c(Y(w_d), \mathcal{O})b_D$ induces an action of the quotient algebra $\mathcal{O}Hc$ and the resulting object is a splendid Rickard complex:

Conjecture 2.21. There is a complex $C \in K^b((\mathcal{O}Gb) \otimes (\mathcal{O}Hb_D)^{\mathrm{opp}}\text{-lperm})$, unique up to isomorphism, with the following properties:

- There is a surjective morphism $f\colon \mathcal{O}\boldsymbol{T}_0^{w_d F} \rtimes B_d \to \mathcal{O}Hb_D$ extending the inclusion $\boldsymbol{T}_0^{w_d F} \subset H$ such that
  - f^*C and $\tilde{R}\Gamma_c(Y(w_d), \mathcal{O})b_D$ are isomorphic in $D^b((\mathcal{O}\boldsymbol{T}_0^{w_d F} \rtimes B_d) \otimes (\mathcal{O}H)^{\mathrm{opp}})$,
  - the map $kB_d \to kC_W(w_d)$ deduced from f by applying $k \otimes_{\mathcal{O}\boldsymbol{T}_0^{w_d F}} -$ is the canonical map.
- C is isomorphic to $\tilde{R}\Gamma_c(Y(w_d), \mathcal{O})b_D$ in $K^b((\mathcal{O}G) \otimes (\mathcal{O}C_G(D))^{\mathrm{opp}}\text{-lperm})$.

Furthermore, such a complex C induces a Rickard equivalence between $\mathcal{O}Gb$ and $\mathcal{O}Hb_D$.

The most crucial and difficult part in that conjecture is to show that we have no non-zero shifted endomorphisms of the complex ("disjunction property"), either for the action of G or for that H.

Conjecture 2.21 is known to hold when $l(w_d) = 1$ [87] and for GL_n and $d = n$ [12]. In the first case, we use good properties of cohomology of curves and prove disjunction for the action of G. In the second case, we study the variety $D(\boldsymbol{U}_0)^F \backslash Y(w_d)$ and prove disjunction for the action of H. This works only for GL_n, for we rely on the fact that induced Gelfand–Graev representations generate the category of projective modules.

Remark 2.22. When $\ell \mid (q-1)$ (case $d = 1$), one can formulate a version of Conjecture 2.21 using the variety $Y(\pi)$ [22, Conjectures 2.15].

Remark 2.23. The version "over K" of Conjecture 2.21 is open, even after restricting to unipotent representations (= applying the functor $K \otimes_{KT^{w_d F}} -$). The action of KB_d on $H_c^*(Y(w_d), K)$ should factor through an action of the Hecke algebra of

$C_W(w_d)$, for certain parameters. This is known in some cases: for $d = 1$ [22], [39], when $d = 2$ (work of Lusztig [67] and joint work with Digne and Michel [39]) and in some other cases [38]. The disjunction property is known for w_d a Coxeter element [66], for GL_n and $d = n - 1$ [38] and in most rank 2 groups [39].

2.5. Local representation theory as non-commutative birational geometry. It is expected that birational Calabi–Yau varieties should have equivalent derived categories (cf. [15]). We view Question 2.6 as a non-commutative version: one can expect that "sufficiently nice" Calabi–Yau triangulated categories are determined by (not too small) quotients. We explain here how this analogy can be made precise, in the setting of McKay's correspondence, via Koszul duality.

2.5.1. 2-elementary abelian defect groups. Let P be an elementary abelian 2-group. Let k be a field of characteristic 2 and $V = P \otimes_{\mathbb{F}_2} k$. Let E be a group of odd order of automorphisms of P. The algebras $kP \rtimes E$ and $\Lambda(V) \rtimes E$ are isomorphic.

Koszul duality (cf. eg [53]) gives an equivalence

$$D^b((\Lambda(V) \rtimes E)\text{-modgr}) \xrightarrow{\sim} D^b_{E \times \boldsymbol{G}_m}(V).$$

2.5.2. McKay's correspondence. Let V be a finite-dimensional vector space over k and E a finite subgroup of $\mathrm{GL}(V)$ of order invertible in k. Recall the following conjecture (independence of the crepant resolution):

Conjecture 2.24 (McKay's correspondence). If $X \to V/E$ is a crepant resolution, then $D^b(X) \simeq D^b_E(V)$.

The conjecture is known to hold when $\dim V = 3$ [16], [14] (in dimension 3, the Hilbert scheme of E-clusters on V is a crepant resolution). It is also known when V is a symplectic vector space and E respects the symplectic structure [9]. See [15, §2.2] for more details.

Examples in dimension > 3 where $E - \mathrm{Hilb}\, V$ is smooth are rare. An infinite family of examples is provided by the following theorem of Sebestean [95]:

Theorem 2.25. *Let $n \geq 2$, let k be a field containing a primitive $(2^n - 1)$-th root of unity ζ and let E be the subgroup of $\mathrm{SL}_n(k)$ generated by the diagonal matrix with entries $(\zeta, \zeta^2, \ldots, \zeta^{2^{n-1}})$. Assume $2^n - 1$ is invertible in k.*

Then $E - \mathrm{Hilb}(\boldsymbol{A}^n_k)$ is a smooth crepant resolution of $\boldsymbol{A}^n_k/E$ and there is an equivalence $D^b_E(\boldsymbol{A}^n_k) \xrightarrow{\sim} D^b(E - \mathrm{Hilb}(\boldsymbol{A}^n_k))$.

The diagonal action of $\boldsymbol{G}_m$ on $\boldsymbol{A}^n_k$ induces an action on $E - \mathrm{Hilb}(\boldsymbol{A}^n_k)$ and the equivalence is equivariant for these actions.

Let $G = \mathrm{SL}_2(2^n)$, let P be the subgroup of strict upper triangular matrices (a Sylow 2-subgroup), and let E be the subgroup of diagonal matrices. The action of E on $P \otimes_{\mathbb{F}_2} \overline{\mathbb{F}}_2$ coincides with the one in Theorem 2.25. Combining the solution of

Conjecture 2.2 for G (Okuyama, [70]) and § 3.2.2, the Koszul duality equivalence, and Theorem 2.25, we deduce a geometric realization of modular representations of $\mathrm{SL}_2(2^n)$ in natural characteristic:

Corollary 2.26. *There is a grading on the principal* 2*-block* A *of* $\overline{\mathbb{F}}_2 G$ *and an equivalence* $D^b(A\text{-modgr}) \xrightarrow{\sim} D^b_{\boldsymbol{G}_m}(E - \mathrm{Hilb}\, \boldsymbol{A}^n_k)$.

Remark 2.27. It should be interesting to study homotopy categories of sheaves on singular varieties and their relation to derived categories of crepant resolutions.

2.6. Perverse Morita equivalences. In this part, we shall describe joint work with J. Chuang [30].

2.6.1. Definitions. Let $\mathcal{A}$, $\mathcal{A}'$ be two abelian categories. We assume every object has a finite composition series. Let $\mathcal{S}$ (resp. $\mathcal{S}'$) be the set of isomorphism classes of simple objects of $\mathcal{A}$ (resp. $\mathcal{A}'$).

Definition 2.28. An equivalence $F\colon D^b(\mathcal{A}) \xrightarrow{\sim} D^b(\mathcal{A}')$ is perverse if there is

- a filtration $\emptyset = \mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots \subset \mathcal{S}_r = \mathcal{S}$,
- a filtration $\emptyset = \mathcal{S}'_0 \subset \mathcal{S}'_1 \subset \cdots \subset \mathcal{S}'_r = \mathcal{S}'$,
- and a function $p\colon \{1, \ldots, r\} \to \mathbb{Z}$,

such that

- F restricts to equivalences $D^b_{\mathcal{A}_i}(\mathcal{A}) \xrightarrow{\sim} D^b_{\mathcal{A}'_i}(\mathcal{A}')$,
- $F[-p(i)]$ induces equivalences $\mathcal{A}_i/\mathcal{A}_{i-1} \xrightarrow{\sim} \mathcal{A}'_i/\mathcal{A}'_{i-1}$.

where $\mathcal{A}_i$ (resp. $\mathcal{A}'_i$) is the Serre subcategory of $\mathcal{A}$ (resp. $\mathcal{A}'$) generated by $\mathcal{S}_i$ (resp. $\mathcal{S}'_i$).

An important point is that $\mathcal{A}'$ is determined, up to equivalence, by $\mathcal{A}$, $\mathcal{S}_\bullet$ and p.

2.6.2. Symmetric algebras. Let A be a symmetric finite dimensional algebra and $\mathcal{A} = A$-mod.

We explain how to construct a perverse equivalence, given any $\mathcal{S}_\bullet$ and p (this cannot be done in general for a nonsymmetric algebra A).

Let I be a subset of $\mathcal{S}$. Given $V \in \mathcal{S}$, let P_V be a projective cover of V, let V_I be the largest quotient of P_V all of which composition factors are in I and let $Q_V \to \ker(P_V \to V_I)$ be a projective cover. We put $T_{A,V}(I) = P_V$ if $V \in S - I$, $T_{A,V}(I) = 0 \to Q_V \to P_V \to 0$ if $V \in I$ (where Q_V is in degree 0) and $T_A(I) = \bigoplus_V T_{A,V}(I)$, a tilting complex.

Let $\mathcal{T}$ be the set of isomorphism classes of families $(T_V)_{V\in S}$, where T_V is an indecomposable bounded complex of finitely generated projective A-modules and $\bigoplus_{V\in S} T_V$ is a tilting complex.

We denote by $\mathcal{P}(\mathcal{S})$ the set of subsets of $\mathcal{S}$. We define an action of $\mathrm{Free}(\mathcal{P}(\mathcal{S})) \rtimes \mathfrak{S}(\mathcal{S})$ on $\mathcal{T}$. The symmetric group acts by permutation of indices and $I \subset \mathcal{S}$ sends $(T_V)_V$ to $(T'_V)_V$ given by $T'_V = F^{-1}(T_{B,V}(I))$, where $B = \mathrm{End}_{D^b(A)}(\bigoplus_V T_V)$ and $F = R\,\mathrm{Hom}^\bullet_A(\bigoplus_V T_V, -)\colon D^b(A) \xrightarrow{\sim} D^b(B)$.

Fix now $\mathcal{S}_\bullet$ a filtration of $\mathcal{S}$ and $p\colon \mathcal{S} \to \mathbb{Z}$. We put

$$(T_V)_V = \mathcal{S}_r^{p(r)} \mathcal{S}_{r-1}^{p(r-1)-p(r)} \dots \mathcal{S}_1^{p(1)-p(2)}((P_V)_V),$$

$T = \bigoplus_V T_V$, $A' = \mathrm{End}_{D^b(A)(T)}$ and $F = R\,\mathrm{Hom}^\bullet_A(T, -)$. Then, F is perverse with respect to $\mathcal{S}_\bullet$ and p.

Remark 2.29. One might ask whether all derived equivalences between finite dimensional symmetric algebras are compositions of perverse equivalences, or at least, if two derived equivalent symmetric algebras can be related by a sequence of perverse equivalences. Many of the derived equivalences in block theory are known to be compositions of perverse equivalences and it would be interesting to see if this is also the case for those of [70].

Remark 2.30. One can expect the equivalences predicted in Conjecture 2.20 will be perverse. The filtration should be provided by Lusztig's a-function.

We expect the action of $\mathrm{Free}(\mathcal{P}(\mathcal{S})) \rtimes \mathfrak{S}(\mathcal{S})$ on $\mathcal{T}$ relates to Bridgeland's space of stability conditions [15, §4].

Remark 2.31. The considerations above are interesting for Calabi–Yau algebras of positive dimension. Given I a subset of $\mathcal{S}$, one obtains a torsion theory that needs not always come from a tilting complex. When $r = 2$ and $|\mathcal{S}_2 - \mathcal{S}_1| = 1$, tilting has been known in string theory as Seiberg duality.

3. Invariants

Invariants of triangulated categories and dg-categories are discussed in [55, §6]. We discuss here some more elementary invariants, used to study finite dimensional algebras.

3.1. Automorphisms of triangulated categories

3.1.1. Rings. Let k be a commutative ring and A be a k-algebra. We denote by $\mathrm{Pic}(A)$ the group of isomorphism classes of invertible (A, A)-bimodules and by $\mathrm{DPic}(A)$ the group of isomorphism classes of invertible objects of the derived category of (A, A)-bimodules: this is the part of the automorphism group of $D(A\text{-Mod})$ that comes from standard equivalences. By Rickard's Theorem, $\mathrm{DPic}(A)$ is invariant under derived equivalences.

The following Proposition has been observed by many people (Rickard, Roggenkamp–Zimmermann, [93, Proposition 3.3], [103, Proposition 3.4],…).

Proposition 3.1. *If A is local, then* $\mathrm{DPic}(A) = \mathrm{Pic}(A) \times \langle A[1] \rangle$.

Given R a flat commutative Z-algebra, there is a canonical morphism $\mathrm{DPic}(A) \to \mathrm{DPic}(A \otimes_Z R)$ (joint work with A. Zimmermann [93, §2.4]). If R is faithfully flat over Z, the kernel of that map is contained in $\mathrm{Pic}(A)$. This is the key point for the following (cf. [103, Proposition 3.5] and [93, Proposition 3.3]):

Theorem 3.2. *Assume A is commutative and indecomposable. Then* $\mathrm{DPic}(A) = \mathrm{Pic}(A) \times \langle A[1] \rangle$.

3.1.2. Invariance of automorphisms. Let A be a finite dimensional algebra over an algebraically closed field k. We denote by $\mathrm{Aut}(A)$ the group of automorphisms of A. This is an algebraic group and we denote by $\mathrm{Inn}(A)$ its closed subgroup of inner automorphisms. We put $\mathrm{Out}(A) = \mathrm{Aut}(A)/\mathrm{Inn}(A)$. We have a morphism of groups $\mathrm{Aut}(A) \to \mathrm{Pic}(A)$, $\alpha \mapsto [A_\alpha]$, where $A_\alpha = A$ as a left A-module and the right action of $a \in A$ is given by right multiplication by $\alpha(a)$. It induces an injective morphism $\mathrm{Out}(A) \to \mathrm{Pic}(A)$.

The following result [91] gives a functorial interpretation of Out, to be compared with the functorial interpretation of $\mathrm{Pic}(X)$ for a smooth projective variety X.

Theorem 3.3. *The functor from the category of affine varieties over k to groups that sends X to the set of isomorphism classes of $(A \otimes A^{\mathrm{opp}} \otimes \mathcal{O}_X)$-modules that are locally free of rank 1 as $(A \otimes \mathcal{O}_X)$ and as $(A^{\mathrm{opp}} \otimes \mathcal{O}_X)$-modules is represented by* $\mathrm{Out}(A)$.

The following theorem [91] shows the invariance of Out^0, the identity component of Out, under certain equivalences. In the case of Morita equivalences, it goes back to Brauer, and for derived equivalences, it has been obtained independently by Huisgen-Zimmermann and Saorín [49]. In these cases, it follows easily from Theorem 3.3 while, for stable equivalences, some work is needed to get rid globally of projective direct summands.

Theorem 3.4. *Let B be a finite dimensional k-algebra and let C be a bounded complex of finitely generated (A, B)-bimodules inducing a derived equivalence or a stable equivalence (in which case we assume A and B are self-injective). Then there is a unique isomorphism of algebraic groups $\sigma : \mathrm{Out}^0(A) \xrightarrow{\sim} \mathrm{Out}^0(B)$ such that $A_\alpha \otimes_A C \simeq C \otimes_B B_{\sigma(\alpha)}$ for all $\alpha \in \mathrm{Out}^0(A)$.*

Yekutieli [104] deduces that $\mathrm{DPic}(A)$ has a structure of a locally algebraic group, with connected component $\mathrm{Out}^0(A)$.

3.1.3. Coherent sheaves. The following result [91] is a variant of Theorem 3.4.

Theorem 3.5. *Let X and Y be two smooth projective schemes over an algebraically closed field k. An equivalence $D^b(X) \xrightarrow{\sim} D^b(Y)$ induces an isomorphism* $\mathrm{Pic}^0(X) \rtimes \mathrm{Aut}^0(X) \xrightarrow{\sim} \mathrm{Pic}^0(Y) \rtimes \mathrm{Aut}^0(Y)$.

This implies in particular that if A and B are derived equivalent abelian varieties, then there is a symplectic isomorphism $\hat{A} \times A \xrightarrow{\sim} \hat{B} \times B$ (and the converse holds as well [71], [74]).

3.1.4. Automorphisms of stable categories and endo-trivial modules. Let A be a finite dimensional self-injective algebra over an algebraically closed field k. We denote by $\mathrm{StPic}(A)$ the group of isomorphism classes of invertible objects of $(A \otimes A^{\mathrm{opp}})$-$\overline{\mathrm{mod}}$.

Let P be an ℓ-group and k a field of characteristic ℓ. A finitely generated kP-module L is an *endo-trivial* module if $L \otimes_k L^* \simeq k$ in kP-$\overline{\mathrm{mod}}$ or equivalently, if $\mathrm{End}_{kP\text{-}\overline{\mathrm{mod}}}(L) = k$ [25]. Note that the classification of endo-trivial modules has been recently completed [27] (the case where P is abelian goes back to [34]).

Let $\mathcal{T}(kP)$ be the group of isomorphism classes of indecomposable endo-trivial modules. We have an injective morphism of groups

$$\mathcal{T}(kP) \to \mathrm{StPic}(kP),\ [L] \mapsto [\mathrm{Ind}_{\Delta P}^{P \times P^{\mathrm{opp}}} L].$$

This extends to an isomorphism $\mathcal{T}(kP) \times \mathrm{Out}(kP) \xrightarrow{\sim} \mathrm{StPic}(kP)$ ([64, §3] and [26, §2]).

Let Q be an ℓ-group. A stable equivalence of Morita type kP-$\overline{\mathrm{mod}} \xrightarrow{\sim} kQ$-$\overline{\mathrm{mod}}$ induces an isomorphism $\mathcal{T}(kP) \xrightarrow{\sim} \mathcal{T}(kQ)$. It actually forces the algebras kP and kQ to be isomorphic ([64, §3], [26, Corollary 2.4]). It is an open question whether this implies that P and Q are isomorphic.

Theorem 3.6 ([26, Theorem 3.2]). *Let P be an abelian ℓ-group and E a cyclic ℓ'-group acting freely on P. We put $G = P \rtimes E$. Then* $\mathrm{StPic}(kG) = \mathrm{Pic}(kG) \cdot \langle \Omega \rangle$. *In particular, the canonical morphism* $\mathrm{TrPic}(kG) \to \mathrm{StPic}(kG)$ *is surjective.*

Remark 3.7. Let A be a block over k of a finite group, with defect group isomorphic to P and $N_G(P)/P$ acting as E on P. From Theorem 3.6, one deduces [26, Corollary 4.4] via a construction of Puig [77], that a stable equivalence of Morita type between A and kG lifts to a Rickard equivalence if and only if A and kG are Rickard equivalent if and only if they are splendidly Rickard equivalent. In particular, for blocks with abelian defect group D such that $N_G(D, b_D)/C_G(D)$ is cyclic, then Conjecture 2.2 implies Conjecture 2.10.

3.2. Gradings. In this section, we describe results of [91].

3.2.1. Transfer of gradings. We assume we are in the situation of Theorem 3.4. Assume A is graded, *i.e.*, there is a morphism $\boldsymbol{G}_m \to \mathrm{Aut}(A)$. The induced morphism $\boldsymbol{G}_m \to \mathrm{Out}^0(A)$ induces a morphism $\boldsymbol{G}_m \to \mathrm{Out}^0(B)$. There exists a lift to a morphism $\boldsymbol{G}_m \to \mathrm{Aut}^0(B)$, and this corresponds to a grading on B. There is a grading on (an object isomorphic to) C that makes it into a complex of graded (A, B)-bimodules and it induces an equivalence between the appropriate graded categories.

Let A be a self-injective indecomposable graded algebra, let n be the largest integer such that $A_n \neq 0$, and let $C \in \mathbb{Z}[q, q^{-1}]$ be the graded Cartan matrix of A.

If A is non-negatively graded and the Cartan matrix of A_0 has non-zero determinant, then $\deg \det(C) = nr$, where r is the number of simple A-modules. As a consequence, one gets a positive solution of a "non-negatively graded" version of Conjecture 2.5:

Proposition 3.8. *Let A and B be two indecomposable self-injective non-negatively graded algebras. Assume A_0 has finite global dimension and there is a graded stable equivalence of Morita type between A and B. Then A and B have the same number of simple modules.*

Remark 3.9. Let A be a non-negatively graded indecomposable self-injective algebra with A_0 of finite global dimension. Let B be a stably equivalent self-injective algebra. One could hope that there is a compatible grading on B that is non-negative, but this is not possible in general. It would be still be very interesting to see if this can be achieved if the grading on A is "tight" in the sense of Cline–Parshall–Scott, *i.e.*, if $\bigoplus_{j \leq i} A_j = (JA)^i$ (cf. the gradings in § 3.2.2).

3.2.2. Blocks with abelian defect. Let P be an abelian ℓ-group and k an algebraically closed field of characteristic ℓ. The algebra kP is (non-canonically) isomorphic to the graded algebra associated to the radical filtration of kP. Fixing such an isomorphism provides a grading on kP. Let E be an ℓ'-group of automorphisms of P. Then the isomorphism above can be made E-equivariant and we obtain a structure of graded algebra on $kP \rtimes E$ extending the grading on kP and with kE in degree 0. Given a central extension of E by $k^\times$, this construction applies as well to the twisted group algebra $k_*P \rtimes E$.

Let A be a block of a finite group over k with defect group D. Then there is E and a central extension as above such that the corresponding block of $N_G(D)$ is Morita equivalent to $k_*D \rtimes E$ [60]. So, Conjecture 2.2 predicts there are interesting gradings on A. In the inductive approach to Conjecture 2.11, there is a stable equivalence of Morita type between A and $k_*D \rtimes E$, and we can provide A with a grading compatible with the equivalence (but we do not know if the grading can be chosen to be non-negative).

Remark 3.10. The gradings on blocks with abelian defect should satisfy some Koszulity properties (cf. [73], as well as work of Chuang). Turner [101] expects that gradings will even exist for blocks of symmetric groups with non abelian defect.

Remark 3.11. Using the equivalences in § 4.1, we obtain gradings on blocks of abelian defect of symmetric groups and on blocks of Hecke algebras over $\mathbb{C}$. One can expect the corresponding graded Cartan matrices to be given in terms of Kazhdan–Lusztig polynomials. So, the equivalences carry some "geometric meaning".

3.3. Dimensions

3.3.1. Definition and bounds. Let us explain how to associate a dimension to a triangulated category $\mathcal{T}$ (cf. [88]). For the derived category of a finite dimensional algebra, this is related to the Loewy length and to the global dimension, none of which are invariant under derived equivalences.

Given $\mathcal{U}_1$ and $\mathcal{U}_2$ two subcategories of $\mathcal{T}$, we denote by $\mathcal{U}_1 * \mathcal{U}_2$ the smallest full subcategory of $\mathcal{T}$ closed under direct summands and containing the objects M such that there is a distinguished triangle

$$M_1 \to M \to M_2 \rightsquigarrow$$

with $M_i \in \mathcal{U}_i$. Given $M \in \mathcal{T}$, we denote by $\langle M \rangle$ the smallest full subcategory of $\mathcal{T}$ containing M and closed under direct summands, direct sums, and shifts. Finally, we put $\langle M \rangle_0 = 0$ and define inductively $\langle M \rangle_i = \langle M \rangle_{i-1} * \langle M \rangle$.

The dimension of $\mathcal{T}$ is defined to be the smallest integer $d \geq 0$ such that there is $M \in \mathcal{T}$ with $\mathcal{T} = \langle M \rangle_{d+1}$ (we set $\dim \mathcal{T} = \infty$ if there is no such d). The notion of finite-dimensionality corresponds to Bondal–Van den Bergh's property of being strongly finitely generated [10].

Given a right coherent ring A, then $\dim D^b(A) \leq \operatorname{gldim} A$ (cf. [59, Proposition 2.6] and [88, Propositions 7.4 and 7.24]).

Let A be a finite dimensional algebra over a field k. Denote by $J(A)$ the Jacobson radical of A. The Loewy length of A is the smallest integer $d \geq 1$ such that $J(A)^d = 0$. We have $\dim D^b(A) < \text{Loewy length}(A)$.

Let X be a separated scheme of finite type over a perfect field k.

Theorem 3.12. *We have* $\dim D^b(X) < \infty$.

- *If X is reduced, then* $\dim D^b(X) \geq \dim X$.
- *If X is smooth and quasi-projective, then* $\dim D^b(X) \leq 2 \dim X$.
- *If X is smooth and affine, then* $\dim D^b(X) = \dim X$.

There does not seem to be any known example of a smooth projective variety X with $\dim D^b(X) > \dim X$, although this is expected to happen, for example when X is an elliptic curve (note nevertheless that $\dim D^b(\boldsymbol{P}^n) = n$).

Note that a triangulated category with finitely many indecomposable objects up to isomorphism has dimension 0. This applies to $D^b(kQ)$, where Q is a quiver of type ADE. This applies also to the orbit categories constructed by Keller (cf. [55, §4.9], [54, §8.4]). They depend on a positive integer d, and they are Calabi–Yau of dimension d.

When $\mathcal{T}$ is compactly generated, the property for $\mathcal{T}^c$ to be finite-dimensional can be viewed as a counterpart of having "finite global dimension".

3.3.2. Representation dimension. Auslander [5] introduced a measure for how far an algebra is from being representation finite. The example of exterior algebras below shows that this notion is pertinent. Let A be a finite dimensional algebra. The representation dimension of A is $\inf\{\mathrm{gldim}(A \oplus A^* \oplus M)\}_{M \in A\text{-mod}}$. This is known to be finite [50].

In [89], we show that this notion is related to the notion of dimension for associated triangulated categories. For example, $\dim D^b(A) \leq \mathrm{repdim}\, A$.

Let A be a non semi-simple self-injective k-algebra. We have

$$2 + \dim A\text{-}\overline{\mathrm{mod}} \leq \mathrm{repdim}\, A \leq \text{Loewy length}(A)$$

(the second inequality comes from [5, §III.5, Proposition]).

The following theorem is obtained by computing $\dim \Lambda(k^n)\text{-}\overline{\mathrm{mod}}$ via Koszul duality. It gives the first examples of algebras with representation dimension > 3.

Theorem 3.13. *Let n be a positive integer. We have* $\mathrm{repdim}\, \Lambda(k^n) = n + 1$.

Remark 3.14. One can actually show more quickly [59] that the algebra with quiver

and relations $x_i x_j = x_j x_i$ has representation dimension $\geq n$, using that its derived category is equivalent to $D^b(\boldsymbol{P}^n)$ [6].

Using the inequality above, one obtains the following theorem, which solves the prime 2 case of a conjecture of Benson.

Theorem 3.15. *Let G be a finite group and k a field of characteristic* 2. *If G has a subgroup isomorphic to $(\mathbb{Z}/2)^n$, then $n <$* Loewy length(kG).

4. Categorifications

This chapter discusses the categorifications of two structures, which are related to derived equivalences. We hope these categorifications will eventually lead to the construction of four-dimensional quantum field theories (as advocated in [33]), via the construction of appropriate tensor structures.

4.1. $\mathfrak{sl}_2$

4.1.1. Abelian defect conjecture for symmetric and general linear groups. Let G be a symmetric group and B an ℓ-block of kG with defect group D. Assume D is abelian and let $w = \log_\ell |D|$. In 1992, a three steps strategy was proposed for Conjecture 2.10 (inspired by the simpler character-theoretic part [84]):

- Rickard equivalence between $k(\mathbb{Z}/\ell \rtimes \mathbb{Z}/(\ell-1)) \wr \mathfrak{S}_w$ and the principal block of $k\mathfrak{S}_\ell \wr \mathfrak{S}_w$;
- Morita equivalence between the principal block of $k\mathfrak{S}_\ell \wr \mathfrak{S}_w$ and B_w;
- Rickard equivalence between B_w and B.

Here, B_w is a certain ℓ-block of symmetric groups (a "good block"). Scopes [94] has constructed a number of Morita equivalences between blocks of symmetric groups. For fixed w, there are only finitely many classes of blocks of symmetric groups up to Scopes equivalence, and B_w is defined to be the largest block that is not Scopes equivalent to a smaller block.

The first equivalence is deduced from an equivalence between the principal blocks of $\mathfrak{S}_\ell$ and $\mathbb{Z}/\ell \rtimes \mathbb{Z}/(\ell-1)$ via Clifford theory [68].

The second equivalence was established by Chuang and Kessar [28], the functor used is a direct summand of the induction functor.

The third equivalence is part of the general problem, raised by Broué, of constructing Rickard equivalences between two blocks of symmetric groups with isomorphic defect groups (equivalently, with same local structure). Rickard [80] constructed complexes of bimodules that he conjectured would solve that problem, generalizing Scopes construction (case where the complex has only one non-zero term). Rickard proved the invertibility of his complexes when they have two non-zero terms. The general case has proven difficult to handle directly.

Remark 4.1. The same strategy applies for general linear groups (in non-describing characteristic). Theorem 2.15 reduces the study to unipotent blocks. Step 2 above was handled in [69], [100]. As pointed out by H. Miyachi, this generalizes Puig's result [76] ($\mathrm{GL}_n(q)$, $\ell \mid (q-1)$).

Remark 4.2. "Good" blocks of symmetric groups have "good" properties. After the Morita equivalence theorem of [28], their properties were first analyzed by Miyachi [69], in the more complicated case of general linear groups: decomposition matrices and radical series of Specht modules were determined in the abelian defect case, by a direct analysis of the wreath product. As a consequence, decomposition matrices were known for good blocks of Hecke algebras in characteristic zero. For good blocks of symmetric groups with abelian defect, as well as for Hecke algebras in characteristic zero, a direct computation of the decomposition numbers is given in [52] (cf. also [51] for earlier results in that direction) and another approach is the determination of the relevant part of the canonical/global crystal basis [31], [32], [61].

For blocks of symmetric groups with non abelian defect, the decomposition matrices can be described in terms of decomposition matrices of smaller symmetric groups and remarkable structural properties are conjectured by Turner [101], [102], [72]. Good blocks have also been used by Fayers for the classification of irreducible Specht modules [40] and to show that blocks of weight 3 have decomposition numbers 0 or 1 (for $\ell > 3$) [41].

4.1.2. Fock spaces. Let us recall the Lie algebra setting for symmetric group representations (cf. e.g. [4]). Let $M = \bigoplus_{n\geq 0} \mathbb{Q} \otimes_{\mathbb{Z}} K_0(\mathbb{C}\mathfrak{S}_n\text{-mod})$. The complex irreducible representations of the symmetric group $\mathfrak{S}_n$ are parametrized by partitions of n and we obtain a basis of M parametrized by all partitions. We view M as a Fock space, with an action of $\hat{\mathfrak{sl}}_\ell$ and we recall a construction of this action, for the generators e_a and f_a (where $a \in \mathbb{F}_\ell$).

We have a decomposition

$$\mathrm{Res}^{\mathbb{F}_\ell\mathfrak{S}_n}_{\mathbb{F}_\ell\mathfrak{S}_{n-1}} = \bigoplus_{a\in\mathbb{F}_\ell} F_a,$$

where $F_a(M)$ is the generalized a-eigenspace of $X_n = (1,n)+(2,n)+\cdots+(n-1,n)$. Taking classes in K_0 and summing over all n, we obtain endomorphisms f_a of

$$V = \bigoplus_{n\geq 0} \mathbb{Q} \otimes_{\mathbb{Z}} K_0(\mathbb{F}_\ell\mathfrak{S}_n\text{-mod}).$$

Using induction, we obtain similarly endomorphisms e_a (adjoint to the f_a). The decomposition lifts to a decomposition of $\mathrm{Res}^{\mathbb{Z}_\ell\mathfrak{S}_n}_{\mathbb{Z}_\ell\mathfrak{S}_{n-1}}$ and we obtain endomorphisms e_a and f_a of M. The decomposition map $M \to V$ and the Cartan map $\bigoplus_{n\geq 0} \mathbb{Q} \otimes_{\mathbb{Z}} K_0(\mathbb{F}_\ell\mathfrak{S}_n\text{-proj}) \to M$ are morphisms of $\hat{\mathfrak{sl}}_\ell$-modules. The image of the Cartan map is the irreducible highest weight submodule L of M generated by $[\emptyset]$.

Let us note two important properties relating the module structure of V and the modular representation theory of symmetric groups:

- The decomposition of V into weight spaces corresponds to the block decomposition.
- Two blocks have isomorphic defect groups if and only if they are in the same orbit under the adjoint action of the affine Weyl group $\tilde{A}_{\ell-1}$.

In order to prove that two blocks of symmetric groups with isomorphic defect groups are derived equivalent, it is enough to consider a block and its image by a simple reflection s_a of $\tilde{A}_{\ell-1}$ (this involves only the $\mathfrak{sl}_2$-subalgebra generated by e_a and f_a). This is the situation in which Rickard constructed his complexes Θ_a.

Remark 4.3. These constructions extend to Hecke algebras of symmetric groups over $\mathbb{C}$, at an ℓ-th root of unity (here, $\ell \geq 2$ can be an arbitrary integer). In that situation, the classes of the indecomposable projective modules form the canonical/global crystal basis of L (Lascoux–Leclerc–Thibon's conjecture, proven by Ariki [3], cf. also [43]).

4.1.3. $\mathfrak{sl}_2$-categorifications. We describe here joint work with J. Chuang [29] (cf. also [90] for a survey and [44], [45], [7], [42] for related work). This is the special case of a more general theory under construction for Kac–Moody algebras.

Let k be an algebraically closed field and $\mathcal{A}$ a k-linear abelian category all of whose objects have finite composition series.

An $\mathfrak{sl}_2$-categorification on $\mathcal{A}$ is the data of

- (E, F) a pair of adjoint exact functors $\mathcal{A} \to \mathcal{A}$,
- $X \in \mathrm{End}(E)$, $T \in \mathrm{End}(E^2)$, $q \in k^\times$, and $a \in k$ (with $a \neq 0$ if $q \neq 1$)

satisfying the following properties:

- $[E]$ and $[F]$ give rise to a locally finite representation of $\mathfrak{sl}_2$ on $K_0(\mathcal{A})$,
- for S a simple object of $\mathcal{A}$, $[S]$ is a weight vector,
- F is isomorphic to a left adjoint of E,
- $(T\mathbf{1}_E) \circ (\mathbf{1}_E T) \circ (T\mathbf{1}_E) = (\mathbf{1}_E T) \circ (T\mathbf{1}_E) \circ (\mathbf{1}_E T)$,
- $(T + \mathbf{1}_{E^2}) \circ (T - q\mathbf{1}_{E^2}) = 0$,
- $T \circ (\mathbf{1}_E X) \circ T = \begin{cases} q(X\mathbf{1}_E) & \text{if } q \neq 1, \\ X\mathbf{1}_E - T & \text{if } q = 1, \end{cases}$
- $X - a\mathbf{1}_E$ is locally nilpotent.

From that data, we define two truncated powers $E^{(n,\pm)}$ (non-canonically isomorphic), using an affine Hecke algebra action on E^n. Following Rickard, we construct a complex Θ with terms $E^{(i,-)}F^{(j,+)}$.

The following theorem is proved by reduction to the case of "minimal categorifications", which are naturally associated to simple representations of $\mathfrak{sl}_2$.

Theorem 4.4. *Θ gives rise to self-equivalences of $K^b(\mathcal{A})$ and $D^b(\mathcal{A})$. This categorifies the action of $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ on $K_0(\mathcal{A})$.*

Remark 4.5. The self-equivalence Θ is perverse (cf. § 2.6), and this is a crucial point in the proof.

The construction of § 4.1.2 provides a structure of $\mathfrak{sl}_2$-categorification on $\mathcal{A} = \bigoplus_{n\geq 0} \overline{\mathbb{F}}_\ell \mathfrak{S}_n$-mod (for a given $a \in \mathbb{F}_\ell$). From the previous theorem, we deduce

Corollary 4.6. *Two blocks of symmetric groups with isomorphic defect groups are splendidly Rickard equivalent.*

Conjecture 2.10 *holds for blocks of symmetric groups.*

This corollary has a counterpart for $\mathrm{GL}_n(\mathbb{F}_q)$ and $\ell \nmid q$.

Remark 4.7. In general, there is a decomposition $\mathcal{A} = \bigoplus_\lambda \mathcal{A}_\lambda$ coming from the weight space decomposition of $K_0(\mathcal{A})$. There is a categorification of $[e, f] = h$ in the form of isomorphisms $EF_{|\mathcal{A}_\lambda} \xrightarrow{\sim} FE_{|\mathcal{A}_\lambda} \oplus \mathrm{Id}_{\mathcal{A}_\lambda}^{\oplus \lambda}$ (for $\lambda \geq 0$).

Remark 4.8. One can give a definition of $\mathfrak{sl}_2$-categorifications for triangulated categories and the definition above becomes a theorem that says that there is an induced categorification on $K^b(\mathcal{A})$ (and on $D^b(\mathcal{A})$).

Remark 4.9. One can also construct $\mathfrak{sl}_2$-categorifications on category $\mathcal{O}$ for $\mathfrak{gl}_n(\mathbb{C})$ and for rational representations of $\mathrm{GL}_n(\overline{\mathbb{F}}_p)$. One deduces from Theorem 4.4 that blocks with the same stabilizers under the affine Weyl groups are derived equivalent (a conjecture of Rickard).

Remark 4.10. The endomorphism X has different incarnations: Jucys–Murphy element, Casimir,….

Remark 4.11. It is expected that the functors Θ_a constructed for $a \in \mathbb{F}_\ell$ provide an action of the affine braid group $B_{\tilde{A}_{\ell-1}}$ on $\bigoplus_n D^b(\mathbb{F}_\ell \mathfrak{S}_n)$.

4.2. Braid groups

4.2.1. Definition. We present here a categorification of braid groups associated to Coxeter groups, following [90]. This should be useful for the study of categories of representations of semi-simple Lie algebras, affine Lie algebras, simple algebraic groups over an algebraically closed field,... On the other hand, work of Khovanov [56] shows its relevance for invariants of links (type A), cf. also [98].

Let (W, S) be a Coxeter group, with S finite. Let V be its reflection representation over $\mathbb{C}$ and let B_W be the braid group of W. Let $A = \mathbb{C}[V]$. Given $s \in S$, let $F_s = 0 \to A \otimes_{A^s} A \xrightarrow{\text{mult}} A \to 0$, where A is in degree 1. This is an invertible object of $K^b(A \otimes A)$. Given two decompositions of an element of B_W in a product of the generators and their inverses, we construct a canonical isomorphism between the corresponding products of F_s. The system of isomorphism coming from the various decompositions of an element $b \in B_W$ is transitive and, taking its limit, we obtain an element $F_b \in K^b(A \otimes A)$. The full subcategory of $K^b(A \otimes A)$ with objects the F_b's defines a strict monoidal category $\mathcal{B}_W$.

We expect that there is a simple presentation of $\mathcal{B}_W$ by generator and relations (or rather of a related 2-category involving subsets of S). This should be related to the vanishing of certain Hom-spaces, for example $\mathrm{Hom}_{K^b(A\otimes A)}(F_b, F_{b'}^{-1}[i])$ should be 0 when b and b' are the canonical lifts of distinct elements of W.

Remark 4.12. The bimodules obtained by tensoring the $A \otimes_{A^s} A$ are Soergel's bimodules. Soergel showed they categorify the Hecke algebra of W. He also conjectured that the indecomposable objects correspond to the Kazhdan–Lusztig basis of W [96], [97].

Remark 4.13. When W is finite, one can expect that there is a construction of $\mathcal{B}_W$ that does not depend on the choice of S. Such a construction might then make sense for complex reflection groups.

4.2.2. Representations and geometry. Let $\mathfrak{g}$ be a complex semi-simple Lie algebra with Weyl group W and let $\mathcal{O}_0$ be the principal block of its category $\mathcal{O}$. It has been widely noticed that there is a weak action of B_W on $D^b(\mathcal{O})$, using wall-crossing

functors. We show that there is a genuine action of B_W on $D^b(\mathcal{O}_0)$ and there is a much more precise statement: there is a monoidal functor from $\mathcal{B}_W$ to the category of self-equivalences of $D^b(\mathcal{O}_0)$. This has a counterpart for the derived category of B-equivariant sheaves on the flag variety (in which case the genuine action of the braid group goes back to [36]). These actions are compatible with Beilinson–Bernstein's equivalence. Conversely, a suitable presentation of $\mathcal{B}_W$ by generators and relations should provide a quick proof of that equivalence (and of affine counterparts), in the spirit of Soergel's construction. The representation-theoretic and the geometrical categories should be viewed as two realizations of the same "2-representation" of $\mathcal{B}_W$. Also, this approach should give a new proof of the results of [2] comparing quantum groups at roots of unity and algebraic groups in characteristic p.

References

[1] Alperin, J., Weights for finite groups. In *The Arcata conference on representations of finite groups*, Vol. 1, Amer. Math. Soc., Providence, RI, 1987, 369–379.

[2] Andersen, H. H., Jantzen, J. C., and Soergel, W., Representations of quantum groups at a pth root of unity and of semisimple groups in characteristic p: independence of p. *Astérisque* **220** (1994).

[3] Ariki, S., On the decomposition numbers of the Hecke algebra of $G(m, 1, n)$. *J. Math. Kyoto Univ.* **36** (4) (1996), 789–808.

[4] Ariki, S., *Representations of quantum algebras and combinatorics of Young tableaux*. Univ. Lecture Ser. 26, American Math. Soc., Providence, RI, 2002.

[5] Auslander, M., *Representation dimension of Artin algebras*. Queen Mary College Mathematics Notes, Queen Mary College, London 1971.

[6] Beilinson, A. A., Coherent sheaves on $\boldsymbol{P}^n$ and problems of linear algebra. *Funct. Anal. Appl.* **12** (1978), 214–216.

[7] Bernstein, J., Frenkel I., and Khovanov, M., A categorification of the Temperley-Lieb algebra and Schur quotients of $U(\mathfrak{sl}_2)$ via projective and Zuckerman functors. *Selecta Math. (N.S.)* **5** (1999), 199–241.

[8] Bessis, D., Digne, F., and Michel, J., Springer theory in braid groups and the Birman-Ko-Lee monoid. *Pacific J. Math.* **205** (2002), 287–309.

[9] Bezrukavnikov, R., and Kaledin, D., McKay equivalence for symplectic resolutions of quotient singularities. *Proc. Steklov Inst. Math.* **246** (2004), 13–33.

[10] Bondal, A., and Van den Bergh, M., Generators and representability of functors in commutative and noncommutative geometry. *Moscow Math. J.* **3** (2003), 1–36.

[11] Bonnafé, C., and Rouquier, R., Catégories dérivées et variétés de Deligne-Lusztig. *Inst. Hautes Études Sci. Publ. Math.* **97** (2003), 1–59.

[12] Bonnafé, C., and Rouquier, R., Coxeter orbits and modular representations. *Nagoya Math. J.*, to appear; math.RT/0511737(v2).

[13] Bonnafé, C., and Rouquier, R., Catégories dérivées et variétés de Deligne-Lusztig, II. In preparation.

[14] Bridgeland, T., Flops and derived categories. *Invent. Math.* **147** (2002), 613–632.

[15] Bridgeland, T., Derived categories of coherent sheaves. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 563–582.

[16] Bridgeland, T., King, A., and Reid, M., The McKay correspondence as an equivalence of derived categories. *J. Amer. Math. Soc.* **14** (2001), 535–554.

[17] Broué, M., Isométries parfaites, types de blocs, catégories dérivées. *Astérisque* **181-182** (1990), 61–92.

[18] Broué, M., Isométries de caractères et équivalences de Morita ou dérivées. *Inst. Hautes Études Sci. Publ. Math.* **71** (1990), 45–63.

[19] Broué, M., Reflection Groups, Braid Groups, Hecke Algebras, Finite Reductive Groups. In *Current Developments in Mathematics, 2000*, International Press, Somerville, MA, 2001, 1–103.

[20] Broué, M., and Malle, G., Zyklotomische Heckealgebren. *Astérisque* **212** (1993), 119–189.

[21] Broué, M., and Michel, J., Blocs et séries de Lusztig dans un groupe réductif fini. *J. Reine Angew. Math.* **395** (1989), 56–67.

[22] Broué, M., and Michel, J., Sur certains éléments réguliers des groupes de Weyl et les variétés de Deligne-Lusztig associées. In *Finite reductive groups*, Progr. Math. 141, Birkhäuser, Boston, MA, 1997, 73–139.

[23] Cabanes, M., and Enguehard, M., *Representation theory of finite reductive groups*. New Math. Monogr.1, Cambridge University Press, Cambridge 2004.

[24] Cabanes, M., and Rickard, J., Alvis-Curtis duality as an equivalence of derived categories. In *Modular representation theory of finite groups*, Walter de Gruyter, Berlin 2001, 157–174.

[25] Carlson, J. F., A characterization of endotrivial modules over p-groups. *Manuscripta Math.* **97** (1998), 303–307.

[26] Carlson, J. F., and Rouquier, R., Self-equivalences of stable modules categories. *Math. Z.* 233 (2000), 165–178.

[27] Carlson, J. F., and Thévenaz, J., The classification of endo-trivial modules. *Invent. Math.* **158** (2004), 389–411.

[28] Chuang, J., and Kessar, R., Symmetric groups, wreath products, Morita equivalences, and Broué's abelian defect group conjecture. *Bull. London Math. Soc.* **34** (2002), 174–184.

[29] Chuang, J., and Rouquier, R., Derived equivalences for symmetric groups and $\mathfrak{sl}_2$-categorification. *Ann. of Math.*, to appear.

[30] Chuang, J., and Rouquier, R., Calabi-Yau algebras and perverse Morita equivalences. In preparation.

[31] Chuang, J., and Tan, K. M., Some canonical basis vectors in the basic $U_q(\widehat{\mathfrak{sl}}_n)$-module. *J. Algebra* **248** (2002), 765–779.

[32] Chuang, J., and Tan, K. M., Filtrations in Rouquier blocks of symmetric groups and Schur algebras. *Proc. London Math. Soc.* **86** (2003), 685–706.

[33] Crane, L., and Frenkel, I. B., Four-dimensional topological quantum field theory, Hopf categories, and the canonical bases. *J. Math. Phys.* **35** (1994), 5136–5154.

[34] Dade, E. C., Endo-permutation modules over p-groups II. *Ann. of Math.* 108 (1978), 317-346.

[35] Dade, E. C., A correspondence of characters. In *The Santa Cruz Conference on Finite Groups* (Santa Cruz, Calif., 1979), Proc. Symp. Pure Math. 37, Amer. Math. Soc., Providence, RI, 1980, 401–403.

[36] Deligne, P., Action du groupe des tresses sur une catégorie. *Invent. Math.* **128** (1997), 159–175.

[37] Deligne, P., and Lusztig, G., Representations of reductive groups over finite fields. *Ann. of Math.* **103** (1976), 103–161.

[38] Digne, F., and Michel, J., Endomorphisms of Deligne-Lusztig varieties. Preprint; math.RT/0509011, *Nagoya Math. J.*, to appear.

[39] Digne, F., Michel, J., and Rouquier, R., Cohomologie des variétés de Deligne-Lusztig. *Adv. Math.*, to appear; math.RT/0410454.

[40] Fayers, M., Irreducible Specht modules for Hecke algebras of type A. *Adv. Math.* **193** (2005), 438–452.

[41] Fayers, M., Decomposition numbers for weight three blocks of symmetric groups and Iwahori-Hecke algebras. *Trans. Amer. Math. Soc.*, to appear.

[42] Frenkel, I., Khovanov, M., and Stroppel, C., A categorification of finite-dimensional irreducible representations of quantum sl(2) and their tensor products. Preprint; math.QA/0511467.

[43] Grojnowski, I., Representations of affine Hecke algebras (and quantum GL_n) at roots of unity. *Internat. Math. Res. Notices* **5** (1994), 215–217.

[44] Grojnowski, I., Affine $\hat{\mathfrak{sl}}_p$ controls the modular representation theory of the symmetric groups and related Hecke algebras. Preprint; math.RT/9907129.

[45] Grojnowski, I., and Vazirani, M., Strong multiplicity one theorems for affine Hecke algebras of type A. *Transform. Groups* **6** (2001), 143–155.

[46] Harris, M. E., Splendid derived equivalences for blocks of finite groups. *J. London Math. Soc.* **60** (1999), 71–82.

[47] Harris, M. E., and Linckelmann, M., Splendid derived equivalences for blocks of finite p-solvable groups. *J. London Math. Soc.* **62** (2000), 85–96.

[48] Hiß, G., On the decomposition numbers of $G_2(q)$. *J. Algebra* **120** (1989), 339–360.

[49] Huisgen-Zimmermann, B., and Saorín, M., Geometry of chain complexes and outer automorphisms under derived equivalence. *Trans. Amer. Math. Soc.* **353** (2001), 4757–4777.

[50] Iyama, O., Finiteness of representation dimension. *Proc. Amer. Math. Soc.* **131** (2003), 1011–1014.

[51] James, G., and Mathas, A., Hecke algebras of type A with $q = -1$. *J. Algebra* **184** (1996), 102–158.

[52] James, G., Lyle, S., and Mathas, A., Rouquier blocks. *Math. Z.* **252** (2006), 511–531.

[53] Keller, B., Deriving DG Categories. *Ann. Sci. École Norm. Sup.* **27** (1994), 63–102.

[54] Keller, B., On triangulated orbit categories. *Doc. Math.* **10** (2005), 551–581.

[55] Keller, B., On differential graded categories. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 151–190.

[56] Khovanov, M., Triply-graded link homology and Hochschild homology of Soergel bimodules. Preprint; math.GT/0510265.

[57] Khovanov, M., Link homology and categorification. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 989–999.

[58] Knörr, R., and Robinson, G. R., Some remarks on a conjecture of Alperin. *J. London Math. Soc.* **39** (1989), 48–60.

[59] Krause, H., and Kussin, D., Rouquier's Theorem on representation dimension. Preprint; math.RT/0505055.v2.

[60] Külshammer, B., Crossed products and blocks with normal defect groups. *Comm. Algebra* **13** (1985) 147–168.

[61] Leclerc, B., and Miyachi, H., Some closed formulas for canonical bases of Fock spaces. *Represent. Theory* **6** (2002), 290–312.

[62] Linckelmann, M., Derived equivalences for cyclic blocks over a p-adic ring. *Math. Z.* **207** (1991), 293–304.

[63] Linckelmann, M., A derived equivalence for blocks with dihedral defect groups. *J. Algebra* **164** (1994), 244–255.

[64] Linckelmann, M., Stable equivalences of Morita type for self-injective algebras and p-groups. *Math. Z.* **223** (1996), 87–100.

[65] Linckelmann, M., On derived equivalences and structure of blocks of finite groups. *Turkish J. Math.* **22** (1998), 93–107.

[66] Lusztig, G., Coxeter Orbits and Eigenspaces of Frobenius. *Invent. Math.* **38** (1976), 101–159.

[67] Lusztig, G., *Representations of Finite Chevalley Groups*. CBMS Reg. Conf. Ser. Math.39, Amer. Math. Soc., Providence, RI, 1978.

[68] Marcus, A., On equivalences between blocks of group algebras: reduction to the simple components. *J. Algebra* **184** (1996), 372–396.

[69] Miyachi, H., Unipotent blocks of finite general linear groups in non-defining characteristic. Ph.D. thesis, Chiba University, 2001.

[70] Okuyama, T., Derived equivalences in $\mathrm{SL}_2(q)$. Preprint, 2000.

[71] Orlov, D., Derived categories of coherent sheaves and equivalences between them. *Russian Math. Surveys* **58** (3) (2003), 511–591.

[72] Paget, R., Induction and decomposition numbers for RoCK blocks. *Q. J. Math.* **56** (2005), 251–262.

[73] Peach, M., Rhombal algebras and derived equivalences. PhD thesis, Bristol, 2004.

[74] Polishchuk, A., *Abelian varieties, theta functions and the Fourier transform*. Cambridge Tracts in Math. 153, Cambridge University Press, Cambridge 2003.

[75] Puig, L., Local block theory in p-solvable groups. In *The Santa Cruz Conference on Finite Groups* (Santa Cruz, Calif., 1979), Proc. Symp. Pure Math. 37, Amer. Math. Soc., Providence, RI, 1980, 385–388.

[76] Puig, L., Algèbres de source de certains blocs des groupes de Chevalley. *Astérisque* **181-182** (1990), 221–236.

[77] Puig, L., Une correspondance de modules pour les blocs à groupes de défaut abéliens. *Geom. Dedicata* **37** (1991), 9–43.

[78] Puig, L., *On the local structure of Morita and Rickard equivalences between Brauer blocks*. Progr. Math. 178, Birkhäuser, Basel 1999.

[79] Rickard, J., Derived categories and stable equivalence. *J. Pure Appl. Algebra* **61** (1989), 303–317.

[80] Rickard, J., Talk at MSRI, Berkeley, 6 November 1990.

[81] Rickard, J., Finite group actions and étale cohomology. *Inst. Hautes Études Sci. Publ. Math.* **80** (1994), 81–94.

[82] Rickard, J., Splendid equivalences: derived categories and permutation modules. *Proc. London Math. Soc.* **72** (1996), 331–358.

[83] Rickard, J., The abelian defect group conjecture. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 121–128.

[84] Rouquier, R., Isométries parfaites dans les blocs à défaut abélien des groupes symétriques et sporadiques. *J. Algebra* **168** (1994), 648–694.

[85] Rouquier, R., The derived category of blocks with cyclic defect groups. In *Derived equivalences for group rings*, Lecture Notes in Math. 1685, Springer-Verlag, Berlin 1998, 199–220.

[86] Rouquier, R., Block theory via stable and Rickard equivalences. In *Modular representation theory of finite groups*, Walter de Gruyter, Berlin 2001, 101–146.

[87] Rouquier, R., Complexes de chaînes étales et courbes de Deligne-Lusztig. *J. Algebra* **57** (2002), 482–508.

[88] Rouquier, R., Dimensions of triangulated categories. Preprint; math.CT/0310134(v3).

[89] Rouquier, R., Representation dimension of exterior algebras. *Invent. Math.*, to appear.

[90] Rouquier, R., Categorification of $\mathfrak{sl}_2$ and braid groups. In *Proceedings of the conference ICRA XI*, to appear.

[91] Rouquier, R., Automorphismes, graduations et catégories triangulées. In preparation.

[92] Rouquier, R., Local constructions in block theory. In preparation.

[93] Rouquier, R., and Zimmermann, A., Picard groups for derived categories. *J. London Math. Soc.* **87** (2003), 197–225.

[94] Scopes, J., Cartan matrices and Morita equivalence for blocks of the symmetric groups. *J. Algebra* **142** (1991), 441–455.

[95] Sebestean, M., Correspondance de McKay et équivalences dérivées. Thèse, Université Paris 7, 2005.

[96] Soergel, W., The combinatorics of Harish-Chandra bimodules. *J. Reine Angew. Math.* **429** (1992), 49–74.

[97] Soergel, W., Kazhdan-Lusztig-Polynome und unzerlegbare Bimoduln über Polynomringen. *J. Inst. Math. Jussieu*, to appear.

[98] Stroppel, C., TQFT with corners and tilting functors in the Kac-Moody case. Preprint; math.RT/0605103.

[99] Tits, J., Normalisateurs de tores. I. Groupes de Coxeter étendus. *J. Algebra* **4** (1966), 96–116.

[100] Turner, W., Equivalent blocks of finite general linear groups in non-describing characteristic. *J. Algebra* **247** (2002), 244–267.

[101] Turner, W., Rock blocks. Preprint, 2004.

[102] Turner, W., On Seven families of algebras. Preprint, 2005.

[103] Yekutieli, A., Dualizing complexes, Morita equivalence and the derived Picard group of a ring. *J. London Math. Soc.* **60** (1999), 723–746.

[104] Yekutieli, A., The derived Picard group is a locally algebraic group. *Algebr. Represent. Theory* **7** (2004), 53–57.

Department of Pure Mathematics, University of Leeds, Leeds, LS2 9JT, UK
and
Institut de Mathématiques de Jussieu, 2 place Jussieu, 75005 Paris, France
E-mail: rouquier@maths.leeds.ac.uk

Algorithmic and asymptotic properties of groups

Mark Sapir*

Abstract. This is a survey of the recent work in algorithmic and asymptotic properties of groups. I discuss Dehn functions of groups, complexity of the word problem, Higman embeddings, and constructions of finitely presented groups with extreme properties (monsters).

Mathematics Subject Classification (2000). Primary 20F65; Secondary 20F10.

Keywords. Turing machine, S-machine, Dehn function, Higman embedding, conjugacy problem, amenable group.

1. Introduction

Although the theory of infinite groups is very rich and full of powerful results, there are very few results having more influence on group theory and surrounding areas of mathematics (especially geometry and topology) as the following five.

- The Boone–Novikov theorem about existence of finitely presented groups with undecidable word problem [8], [35].
- The Higman theorem about embeddability of recursively presented groups into finitely presented groups [28];
- The Adian–Novikov solution of the Burnside problem [36].
- Gromov's theorem about groups with polynomial growth [23].
- Olshanskii and his students' theorems about existence of groups with all proper subgroups cyclic (Tarski monsters), and other finitely generated groups with extreme properties [38].

In this paper, I am going to survey the last ten years of my work on the topics related to these results.

Acknowledgement. Most of the work surveyed here is joint with J.-C. Birget, C. Druţu, V. Guba, A. Olshanskii and E. Rips. I am very grateful to them for cooperation.

*This work was supported in part by the NSF grants DMS 0245600 and DMS-0455881.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. *S*-machines

Recall that a Turing machine, say, with one tape is a triple (Y, Q, Θ) where Y is a tape alphabet, Q is the set of states, Θ is a set of commands (transitions) of the form $\theta = [U \to V]$ where U has the form vqu and V has the form $v'q'u'$. Here u, v, u' and v' are words in the tape alphabet, $q, q' \in Q$. A *configuration* of the Turing machine is a word wqw' where w, w' are words in the tape letters, q is a state letter. To apply the command $[U \to V]$ to a configuration, one has to replace U by V.

In order to specify a Turing machine with many tapes, one needs several disjoint sets of state letters. A *configuration* of the machine is a word of the form $u_1q_1u_2 \dots u_Nq_Nu_{N+1}$ where q_i are state letters, u_i are words in tape letters. Of course one needs to separate tapes. That can be done by using the special symbols (endmarkers) marking the beginning and the end of each tape. But these symbols can be treated as state letters as well. Every transition has the form $[U_1 \to V_1, \dots, U_N \to V_N]$, where $[U_i \to V_i]$ is a transition of a 1-tape machine.

Among all configurations of a machine M, one chooses one *accept* configuration W. Then a configuration W_1 is called *accepted* if there exists a *computation* $W_1 \to W_2 \to \dots \to W_n = W$ where each step consists in application of a command of W.

Recall that the time function of a (non-deterministic) Turing machine is the smallest function $f(n)$ such that every accepted input w of size at most n requires at most $f(n)$ steps of the machine to be accepted.

The "common denominator" of the proofs of most of the results I am reviewing here is the notion of an S-machine that I introduced in [50]. Roughly speaking, S-machines make building groups with prescribed properties as easy as programming a Turing machine.

Essentially, an S-machine is simply an HNN-extension of a free group, although not every HNN-extension of a free group is an S-machine.

Let us start with an example that we shall call the *Miller machine*. It is the famous group of C. Miller [34]. Let $G = \langle X \mid R \rangle$ be a finitely presented group. The Miller machine is the group $M(G)$ generated by $X \cup \{q\} \cup \{\theta_x \mid x \in X\} \cup \{\theta_r \mid r \in R\}$ subject to the following relations

$$\theta x = x\theta, \quad \theta_x xq = qx\theta_x, \quad \theta_r q = qr\theta_r$$

where θ is any letter in $\Theta = \{\theta_x \mid x \in X\} \cup \{\theta_r \mid r \in R\}$. Clearly, this is an HNN-extension of the free group $\langle X, q\rangle$ with free letters $\theta \in \Theta$. The main feature of $M(G)$ discovered by Miller is that $M(G)$ *has undecidable conjugacy problem provided* G *has undecidable word problem.* In fact it is easy to see that qw is conjugated to q in $M(G)$ if and only if $w = 1$ in G.

To see that $M(G)$ can be viewed as a machine, consider any word uqv where u, v are words in $X \cup X^{-1}$. If we conjugate uqv by θ_r, we get the word $uqrv$ because $\theta_r q = qr\theta_r$ and θ_r commutes with u and v (here and below we do not distinguish words that are freely equal). Hence conjugation by θ_r amounts to executing

a command $[q \to qr]$. Similarly, conjugation by θ_x amounts to executing a command $[q \to x^{-1}qx]$. If u ends with x, then executing this command means moving q one letter to the left. Thus conjugating words of the form uqv by θ's and their inverses, we can move the "head" q to the left and to the right, and insert relations from R.

The work of the Miller machine $M(G)$ can be drawn in the form of a diagram (see Figure 1) that we call a *trapezium*. It is a tessellation of a disc. Each cell corresponds to one of the relations of the group. The bottom layer of cells in Figure 1 corresponds to the conjugation by θ_x, the next layer corresponds to the conjugation by θ_r, etc. These layers are the so-called θ*-bands*. The bottom side of the boundary of the trapezium is labeled by the first word in the computation (uqv), the top side is labeled by the last word in the computation (q), the left and the right sides are labeled by the *history of computation*, the sequence of θ's and their inverses corresponding to the commands used in the *computation* $uqv \to \cdots \to q$. The words written on the top and bottom sizes of the θ-bands are the intermediate words in the computation. We shall always assume that they are freely reduced.

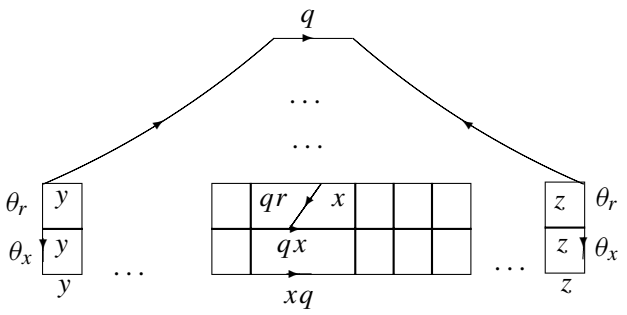


Figure 1. Trapezium of the Miller machine for a deduction $uqv \to \cdots \to q$. Here $u = y \dots x$, $v = \dots z$.

The Miller machine has one tape and one state letter. General S-machines can have many tapes and many state letters. Here is a formal definition.

Let $F(Q, Y)$ be the free group generated by two sets of letters $Q = \bigcup_{i=1}^{N} Q_i$ and $Y = \bigcup_{i=1}^{N-1} Y_i$ where Q_i are disjoint and non-empty (below we always assume that $Q_{N+1} = Q_1$, and $Y_N = Y_0 = \emptyset$).

The set Q is called the set of q-letters, the set Y is called the set of a-letters.

In order to define an HNN-extension, we consider also a collection Θ of N-tuples of θ-letters. Elements of Θ are called *rules*. The components of θ are called *brothers* $\theta_1, \dots, \theta_N$. We always assume that all brothers are different. We set $\theta_{N+1} = \theta_1$, $Y_0 = Y_N = \emptyset$.

With every $\theta \in \Theta$, we associate two sequences of elements in $F(Q \cup Y)$: $B(\theta) = [U_1, \dots, U_N]$, $T(\theta) = [V_1, \dots, V_N]$, and a subset $Y(\theta) = \cup Y_i(\theta)$ of Y, where $Y_i(\theta) \subseteq Y_i$.

The words U_i, V_i satisfy the following restriction:

(*) For every $i = 1, \dots, N$, the words U_i and V_i have the form

$$U_i = v_{i-1} k_i u_i, \quad V_i = v'_{i-1} k'_i u'_i$$

where $k_i, k'_i \in Q_i$, u_i and u'_i are words in the alphabet $Y_i^{\pm 1}$, v_{i-1} and v'_{i-1} are words in the alphabet $Y_{i-1}^{\pm 1}$.

Now we are ready to define an S-machine $\mathcal{S}$ by generators and relations. The generating set X of the S-machine $\mathcal{S}$ consists of all q-, a- and θ-letters. The relations are:

$$U_i \theta_{i+1} = \theta_i V_i, \ i = 1, \dots, s, \quad \theta_j a = a \theta_j$$

for all $a \in Y_j(\theta)$. The first type of relations will be called (q, θ)-*relations*, the second type (a, θ)-*relations*.

Sometimes we will denote the rule θ by $[U_1 \to V_1, \dots, U_N \to V_N]$. This notation contains all the necessary information about the rule except for the sets $Y_i(\theta)$. In most cases it will be clear what these sets are: they are usually equal to either Y_i or $\emptyset$. By default $Y_i(\theta) = Y_i$.

Every S-rule $\theta = [U_1 \to V_1, \dots, U_s \to V_s]$ has an inverse $\theta^{-1} = [V_1 \to U_1, \dots, V_s \to U_s]$; we set $Y_i(\theta^{-1}) = Y_i(\theta)$.

Remark 2.1. Every S-machine is indeed an HNN-extension of the free group $F(Y, Q)$ with finitely generated associated subgroups. The free letters are θ_1 for every $\theta \in \Theta$. We leave it as an exercise to find the associated subgroups.

Every Turing machine T can be considered as an S-machine $S'(T)$ in the natural way: the generators of the free group are all tape letters and all state letters. The commands of the Turing machine are interpreted as rules of the S-machine. The main problem in that conversion is the following: there is a much bigger freedom in applying S-rules than in executing the corresponding commands of the Turing machine. Indeed, the Turing machine is in general not *symmetric* (i.e. if $[U \to V]$ is a command of the Turing machine then $[V \to U]$ is usually not) while every S-machine is symmetric. Another difference is that Turing machines work only with positive words, and S-machines work with arbitrary group words. Hence the language accepted by $S'(T)$ is usually much bigger than the language accepted by T.

Nevertheless, it can be proved that if T is symmetric, and a computation $w_1 \to w_2 \to \cdots$ of the S-machine $S'(T)$ involves only positive words, then that is a computation of T.

This leads to the following idea of converting any Turing machine T into an S-machine $S(T)$. First we construct a symmetric Turing machine T' that is equivalent to T (recognizes the same language). That is a fairly standard Computer Science trick (see [50]): the machine T' first guesses a computation of T, then executes it,

then erases all the tapes. Note that the time function and the space function of T' are equivalent to the time function of T.

The second step is to compose the S-machine $S'(T')$ with a machine that checks positivity of a word. That machine starts working after every step of $S'(T')$. That is if an application of a rule of $S'(T')$ gives a non-positive (reduced) word then the checking machine does not allow the machine $S'(T')$ to proceed to the next step.

There are several checking machines. One of them – the *adding machine* – is very simple but its time function is exponential (see [42]). Another one is very complicated but it has a quadratic time function (see [50]).

Here is the definition of the adding machine. We present it here also in order to show an example of a program of an S-machine. It is not difficult to program an S-machine, but it does require some practice.

Let A be a finite set of letters. Let the set A_1 be a copy of A. It will be convenient to denote A by A_0. For every letter $a_0 \in A_0$, a_1 denotes its copy in A_1. The set of state letters of the adding machine $Z(A)$ is $P_1 \cup P_2 \cup P_3$ where $P_1 = \{L\}$, $P_2 = \{p(1), p(2), p(3)\}$, $P_3 = \{R\}$. The set of tape letters is $Y_1 \cup Y_2$ where $Y_1 = A_0 \cup A_1$ and $Y_2 = A_0$.

The adding machine $Z(A)$ has the following rules (there a is an arbitrary letter from A) and their inverses. The comments explain the meanings of these rules.

- $r_1(a) = [L \to L,\ p(1) \to a_1^{-1} p(1) a_0,\ R \to R]$.

  Comment. The state letter $p(1)$ moves left searching for a letter from A_0 and replacing letters from A_1 by their copies in A_0.

- $r_{12}(a) = [L \to L,\ p(1) \to a_0^{-1} a_1 p(2),\ R \to R]$.

  Comment. When the first letter a_0 of A_0 is found, it is replaced by a_1, and p turns into $p(2)$.

- $r_2(a) = [L \to L,\ p(2) \to a_0 p(2) a_0^{-1},\ R \to R]$.

  Comment. The state letter $p(2)$ moves toward R.

- $r_{21} = [L \to L,\ p(2) \to p(1),\ R \to R]$, $Y_1(r_{21}) = Y_1$, $Y_2(r_{21}) = \emptyset$.

  Comment. $p(2)$ and R meet, the cycle starts again.

- $r_{13} = [L \to L,\ p(1) \to p(3),\ R \to R]$, $Y_1(r_{13}) = \emptyset$, $Y_2(r_{13}) = A_0$.

  Comment. If $p(1)$ never finds a letter from A_0, the cycle ends, $p(1)$ turns into $p(3)$; p and L must stay next to each other in order for this rule to be executable.

- $r_3(a) = [L \to L,\ p(3) \to a_0 p(3) a_0^{-1},\ R \to R]$, $Y_1(r_3(a)) = Y_2(r_3(a)) = A_0$.

  Comment. The letter $p(3)$ returns to R.

The underlying algorithm of the adding machine is simple: the machine starts with a word $Lwp(1)R$, where w is a word in $A \cup A^{-1}$, L, $p(1)$, R are state letters. It

considers the sequence of indexes of the letters in w as a binary number. The initial number is 0. The machine proceeds by adding 1 to this number until it produces $2^n - 1$ where n is the length of the word (each cycle of the machine adds a 1). After that, the machine returns the word to its initial state (all indexes are 0). If the initial word contained a negative letter, the state letter of the adding machine never becomes $p(3)$.

To compose a checking machine Z with an S-machine $\mathcal{S}$ means inserting state letters of Z between any two consecutive state letters of $\mathcal{S}$, and changing the rules of $\mathcal{S}$ in an appropriate way: every rule of $\mathcal{S}$ "turns on" the checking machines. After they finish their work, $\mathcal{S}$ can apply another rule (provided the word is still positive). If Z is a checking machine then the composition of $\mathcal{S}$ and Z is denoted by $\mathcal{S} \circ Z$.

The following results from [50] are very important for the applications. The equivalence of S-machines, their time functions, space functions, etc. are defined as for ordinary Turing machines.

We say that two increasing functions $f, g\colon \mathbb{Z}_+ \to \mathbb{Z}_+$ are equivalent if

$$\frac{1}{C} g\left(\frac{n}{C}\right) - Cn \le f(n) \le Cg(Cn) + Cn \tag{1}$$

for some constant C. We are not going to distinguish equivalent functions in this note. Thus $n^{3.2}$ is the same as $5n^{3.2}$ but different from $n^{3.2} \log n$.

Theorem 2.2 (Sapir, [50]). *Let T be a Turing machine. Then there exists an S-machine $\mathcal{S}$ that is polynomially equivalent to T. Moreover the time function of $\mathcal{S}$ is equivalent* (*in the sense of* (1)) *to the cube of the time function of T, the space function of $\mathcal{S}$ is equivalent to the time function of T.*

Moreover, one can use Miller's machines to simulate any Turing machine.

Theorem 2.3 (Sapir, [45]). *For every Turing machine T there exists a finitely presented group G such that the Miller machine $M(G)$ is polynomially equivalent to T.*

Thus any Turing machine can be effectively simulated by an S-machine with one tape and only one state letter.

3. Dehn functions and the word problem

3.1. The definition. Let $G = \langle X \mid R \rangle$ be a finitely presented group. We shall always assume that $X = X^{-1}$, R is a collection of words in the alphabet X closed under taking inverses and cyclic shifts, i.e. if $r \in R$, $r \equiv ab$ then $r^{-1} \in R$ and $ba \in R$.

The word problem in G asks, given a word w in X (i.e. a product of generators of G), whether w is equal to 1 in G. Clearly, the word problem in all "ordinary" groups is algorithmically decidable. For example, if the group is linear, then in order to check if $w = 1$, one can just multiply matrices representing the generators of G in the order of their appearance in w.

About 70 years ago, van Kampen noticed that $w = 1$ in G if and only if one can tessellate a disc with boundary labeled by w by tiles (cells) whose boundaries are labeled by words in R.

That tessellation is called a *van Kampen* diagram for w. It is also sometimes called *Dehn* or *disc* diagram of w. For example, the trapezium in Figure 1 is a van Kampen diagram over the S-machine $M(G)$ with boundary label $huqvh^{-1}q^{-1}$ where h is the history of the computation.

For every $w = 1$ in G, the area $\mathrm{a}(w)$ is the smallest number of cells in the van Kampen diagram for w, or the simplicial area of the (null-homotopic) loop labeled by w in the Cayley complex Cayley(G, X). Combinatorially, that is the smallest number of factors in any representation of w as a product of conjugates of the words from R. From the logic point of view, that is the length of the shortest "proof" that $w = 1$ in G (steps of the "proof" are insertions of relations from R into W).

It is easy to see that the word problem in G is decidable if and only if the area of a word w representing 1 in G is bounded from above by a recursive function in the length of w. Madlener and Otto [33] and, independently, Gersten [19] introduced a very basic characteristic of the algorithmic complexity of a group G, the Dehn function $\delta_G(n)$ of G: *it is the smallest function $d(n)$ such that the area of a word w of length $\leq n$ representing* 1 *in G does not exceed $d(n)$*. Of course $\delta_G(n)$ depends on the choice of generating set X. But Dehn functions corresponding to different generating sets are *equivalent* in the sense of (1). Similarly, one can introduce the *isodiametric function* of G by looking at the diameter of van Kampen diagrams instead of their areas.

For example, the area of the trapezium in Figure 1 is approximately $|h|$ times the length of the longest θ-band in that trapezium, that can be interpreted as the product of the time of the computation by its space. That observation is the key to converting properties of the S-machine into the properties of the Dehn function.

3.2. The description. Dehn functions reflect in an easy and natural way both geometric and algorithmic properties of a group, so it is natural to ask which functions appear as Dehn functions of groups.

The first observation is not difficult.

Theorem 3.1 (See [50, Theorem 1.1]). *Every Dehn function of a finitely presented group G is* (*equivalent to*) *the time function of a Turing machine solving* (*non-deterministically*) *the word problem in G.*

The proof of this theorem in [50] is more complicated than it should have been. An easier proof can be obtained by using [45, Lemma 1].

Not every increasing function can be equivalent to a time function of a Turing machine. For example, if a time function $f(n)$ does not exceed a recursive function then it must be recursive. On the other hand, any "natural" function is the time function of a Turing machine. In particular, if $f(n)$ can be computed in time $\leq f(n)$ then $f(n)$ is the time function of a deterministic Turing machine computing $f(n)$.

By a theorem proved by Gromov and Olshanskii among others, every finitely presented group with subquadratic Dehn function is in fact hyperbolic, so its Dehn function is linear. It is possible to deduce from a result of Kapovich and Kleiner [29] that if the Dehn function is subquadratic even on an infinite subset of natural numbers then the group is still hyperbolic. Dehn functions of nilpotent groups are bounded by a polynomial [5]. Another source of groups with polynomial Dehn function is the class of groups with simply connected asymptotic cones (Gromov, [24]). Moreover if the asymptotic cones are simply connected then the isodiametric function of the group is linear.

Recall that the asymptotic cone of a group G (see [24], [16] or [18]) is the ultralimit (or Gromov–Hausdorff limit) of a sequence X/d_i where X is a Cayley graph of G, $\lim d_i = \infty$, X/d_i is the metric space X with distance function divided by d_i [23]. Asymptotic cones capture "global" geometric properties of the group G.

Of a particular interest are groups with quadratic Dehn function. That class of groups includes the classes of automatic groups and CAT(0)-groups. Higher dimensional Heisenberg groups [3], [43] and some solvable non-virtually nilpotent groups [17] also have quadratic Dehn functions. That class contains more complicated groups as well. The most striking example so far is the R. Thompson group

$$F = \langle x_0, x_1 \mid x_1^{x_0^2} = x_1^{x_0 x_1}, x_1^{x_0^3} = x_1^{x_0^2 x_1} \rangle$$

where $a^b = b^{-1}ab$. Recall that F is the group of all piecewise linear increasing self-homeomorphisms of the unit interval with finitely many dyadic singular points and all slopes powers of 2. Guba showed in [25] that F has a quadratic Dehn function. One of the most interesting unsolved problems about this class is whether $\mathrm{SL}_n(\mathbb{Z})$ belongs to it for $n \geq 4$.

A very non-trivial result of Bridson and Groves [11] shows that every cyclic extension of a finitely generated free group has quadratic Dehn function. On the other hand, Olshanskii and I proved [42] that HNN extensions of free groups having undecidable conjugacy problem must have Dehn function at least $n^2 \log n$. Together with the result of Bridson and Groves it gives another proof of decidability of the conjugacy problem for cyclic extensions of free groups [7].

It is still unknown whether every group with quadratic Dehn function has decidable conjugacy problem. Olshanskii and I gave a "quasi-proof" of that in [42].

I think that it is most probable that the class of Dehn functions $\geq n^2 \log n$ is as wide as the class of time functions of Turing machines. The next theorem confirms that conjecture in the case of Dehn functions $\geq n^4$.

Theorem 3.2 (See [50]). 1. *Let $\mathcal{D}_4$ be the set of all Dehn functions $d(n) \geq n^4$ of finitely presented groups. Let $\mathcal{T}_4$ be the set of time functions $t(n) \geq n^4$ of arbitrary Turing machines. Let $\mathcal{T}^4$ be the set of superadditive functions which are fourth powers of time functions. Then*

$$\mathcal{T}^4 \subseteq \mathcal{D}_4 \subseteq \mathcal{T}_4.$$

2. *For every time function $T(n)$ of a non-deterministic Turing machine with superadditive $T^4(n)$ there exists a finitely presented group G with Dehn function $T^4(n)$ and the isodiametric function $T^3(n)$.*

Recall that a function f is superadditive if $f(n+m) \geq f(m) + f(n)$ for any m, n. The question of whether all Dehn functions are superadditive is one of the unsolved mysteries of the subject. Together with Victor Guba [26], we proved that *the Dehn function of any non-trivial free product is superadditive.* Thus if there are non-superadditive Dehn functions then there are groups G such that G and $G * \mathbb{Z}$ have different Dehn functions!

Theorem 3.2 has many corollaries. For example, it implies that the *isoperimetric spectrum*, i.e. the set of α's such that $\lfloor n^\alpha \rfloor$ is a Dehn function, contains all numbers $\alpha \geq 4$ whose n-th digit can be computed by a deterministic Turing machine in time less than 2^{2^n}. All "constructible" numbers (rational numbers, algebraic numbers, values of elementary functions at rational points, etc.) satisfy this condition. On the other hand, Theorem 3.1 implies that if α is in the isoperimetric spectrum then the n-th digit of α can be computed in time $\leq 2^{2^{2^n}}$ (see [50] for details). The difference in the number of 2's in these expressions, is the difference between P and NP in Computer Science (if $P = NP$ then there should be two 2's in both expressions).

Note that before [50] has been submitted to Annals of Mathematics (in 1997), only a discrete set of non-integer numbers in the isoperimetric spectrum was known [10]. By the time the paper appeared in print (2002), that set increased by a dense subset in $[2, \infty)$ [9]. Groups in [10] with Dehn functions $\lfloor n^\alpha \rfloor$, $\alpha \notin \mathbb{N}$, have easier presentations than groups based on S-machines having the same Dehn functions, but the construction in [10], [9] is far from universal, and one cannot expect anything like Theorem 3.2 proved using their methods.

Other applications of Theorem 3.2 are:

- the first example of a finitely presented group with NP-complete word problem,
- examples of finitely presented groups with easy word problem (solvable in quadratic time) and arbitrary large (recursive) Dehn functions.

3.3. The proof. Here is how our construction from [50] works. Take any Turing machine M. Let M' be the symmetric Turing machine described above. Let $S(M')$ be the S-machine obtained as a composition of $S'(M')$ with a positivity checking S-machine from [50] working in quadratic time. The time function of $S(M')$ is T^3 and the space function is T where T is the time function of M. We can assume that the accepting configuration of $S(M')$ is some fixed word W of the form $k_1 w_1 k_2 w_2 \dots k_N$ where $N > 8$ (for some small cancellation reasons) and all w_i are copies of each other written in disjoint alphabets and containing no tape letters. That can be achieved by taking N copies of the initial Turing machine and making all of them work in parallel. Finally add one *hub* relation $W = 1$ to the S-machine $S(M')$. The resulting

group $G(M)$ has Dehn function T^4 provided T^4 is superadditive, and isodiametric function T^3.

The main idea of the proof is the following. Take the standard trapezium corresponding to a computation $W_1 \to \cdots \to W_n = W$, identify its left and right sides (which have the same label). The resulting diagram has one hole with boundary label W. Insert the cell corresponding to the hub relation $W = 1$ into the hole. The result is a van Kampen diagram, called a *disc corresponding to the equality* $W_1 = 1$. The area of that diagram is equal to the area of the trapezium (plus 1). So it is equivalent to the product of the time of the computation by its space. The diameter of the disc with perimeter $\leq n$ is the time of the computation. Hence the worst area we can get is T^4, and the worst diameter is T^3. That gives the lower bound of the Dehn function and the isodiametric function. The upper bound is obtained by using certain surgeries on van Kampen diagrams. It turns out that every van Kampen diagram over the group $G(M)$ can be decomposed into a few discs and a diagram whose area is at most cubic (with respect to the perimeter of the original diagram). Thus if the area of a van Kampen diagram is large then most of the area is concentrated in the discs. It turns out also that the sum of the perimeters of the discs does not exceed a constant multiple of the perimeter of the diagram. This gives the desired upper bound of T^4 for the Dehn function (it is in this part of the proof where the superadditivity of T^4 is used) and T^3 for the isodiametric function.

3.4. The Dehn functions of *S*-machines and chord diagrams. It is easy to see that the Dehn function of an S-machine is at most cubic. Indeed, every van Kampen diagram with perimeter of length n over the presentation of an S-machine is covered by θ-bands that start and end on the boundary. There are also q-bands composed of (q, θ)-cells, and a-bands composed of the commutativity (a, θ)-cells. It can be proved that every θ-band intersects a q-band (an a-band) at most once. Hence the total number of (q, θ)-cells is at most n^2. Every other cell is an (a, θ)-cell. Each a-band starts on the boundary of the diagram or on the boundary of a (q, θ)-cell, so the total number of such bands is at most n^2 and the length of each of them is at most n. Hence the total area is at most n^3.

It was conjectured by Rips and myself that the Dehn function of an S-machine should in fact depend on the program of the S-machine. We thought that S-machines should provide examples of groups with Dehn functions strictly between n^2 and n^3. It turned out to be the case. In particular, Olshanskii and I proved in [42] that if $\mathcal{A}$ is any S-machine accepting language L, then the composition $\mathcal{A} \circ Z(A)$ of $\mathcal{A}$ and the adding machine has Dehn function at most $n^2 \log n$ and accepts the same language L. Hence we get the following result.

Theorem 3.3 ([42]). *There exists an S-machine with undecidable conjugacy problem and Dehn function* $n^2 \log n$.

The idea of analyzing van Kampen diagrams over S-machines is to show that if the area of a diagram is large then "most" of the area is inside large subtrapezia, and then

analyze trapezia (i.e. computations of the S-machines). Note that in a trapezium, every θ-band intersects every q-band, thus the band structure of a trapezium is somewhat regular. In order to analyze irregular diagrams, Olshanskii introduced a measure of irregularity, the *dispersion*. In fact, the dispersion is an invariant of the *cord diagram* associated with every van Kampen diagram over an S-machine: the role of chords is played by the θ-bands (T-chords) and the q-bands (Q-chords). It is similar to a Vassiliev invariant of knots.

As I mentioned before, $n^2 \log n$ is the smallest Dehn function of an HNN extension of a free group with undecidable conjugacy problem. If the undecidability condition is dropped, one can construct Dehn functions strictly between n^2 and $n^2 \log n$. In particular, Olshanskii [40] constructed an *S-machine with non-quadratic Dehn function bounded from above by a quadratic function on arbitrary long intervals.* This gives the first example of a finitely presented group with two non-homeomorphic asymptotic cones [41]: *one of the asymptotic cones of this group is simply connected, and another one is not.*

Non-finitely presented groups with "very many" asymptotic cones are constructed in [18] using completely different methods.

Theorem 3.4 (Druţu, Sapir [18]). *There exist finitely generated groups with continuously many* (*maximal theoretically possible if the Continuum Hypothesis is true* [32]) *non-homeomorphic asymptotic cones.*

It is very interesting whether one can replace "finitely generated" by "finitely presented" in Theorem 3.4. One can try to use Higman embeddings from Section 4 to construct such examples.

3.5. Non-simply connected asymptotic cones. Note that Theorem 3.2 gave some of the first examples of groups with polynomial Dehn function and non-simply connected asymptotic cones because their Dehn functions can be polynomial (if the original Turing machine had polynomial time function) while their isodiametric functions are not linear. The first examples of groups with polynomial (cubic) Dehn functions, linear isodiametric functions and non-simply connected asymptotic cones were given in [44]. That answered a question of Druţu from [16].

The groups in [44] are S-machines. The easiest example is this:

$$G = \langle \theta_1, \theta_2, a, k \mid a^{\theta_i} = a, k^{\theta_i} = ka, i = 1, 2 \rangle.$$

The S-machine has one tape letter, one state letter and two rules $[k \to ka]$ (and their inverses).

There is also an S-machine with Dehn function $n^2 \log n$ satisfying the same asymptotic properties. Note that $n^2 \log n$ cannot be lowered to n^2 because of a result of Papasoglu [49]: *all groups with quadratic Dehn functions have simply connected asymptotic cones.*

If a group has non-simply connected asymptotic cone, it is natural to ask what is its fundamental group. We do not know what are the fundamental groups of asymptotic

cones of S-machines. These groups may provide some interesting invariants of S-machines and Turing machines, so it is worthwhile studying them.

The following theorem gives a partial answer to the question of what kind of groups can be fundamental groups of asymptotic cones of finitely generated groups.

Theorem 3.5 ([18]). *For every countable group C there exists an asymptotic cone of a finitely generated group G whose fundamental group is isomorphic to the free product of continuously many copies of C.*

The proof does not use S-machines but uses some small cancellation arguments. It would be interesting to find finitely presented groups with similar "arbitrary" fundamental groups of asymptotic cones. Perhaps the Higman embeddings discussed in the next section will help solving that problem. Another very interesting problem (due to Gromov [24]) is whether there exists an asymptotic cone of a finitely generated group with non-trivial but at most countable fundamental group.

4. Higman embeddings

The flexibility of S-machines allowed us to construct several versions of Higman embeddings (embeddings of recursively presented groups into finitely presented ones) preserving certain properties of the group.

4.1. An easy construction. The easiest known construction of a Higman embedding is the following. Let H be a recursively presented group $\langle X \mid R\rangle$. Then the set of all words in $X \cup X^{-1}$ that are equal to 1 in H is recursively enumerable. Hence we can assume that R consists of all these words. Then there exists an S-machine recognizing R. More precisely, for every word w in X, it starts with a word $q_1 w q_2 q_3 \dots q_m$ and ends with a word $\bar q_1 \bar q_2 \dots \bar q_m$ if and only if $w \in R$.

Again, as in the proof of Theorem 3.2, we consider $N > 8$ copies of this S-machine and assume that the input of the S-machine $\mathcal{S}$ has the form

$$K(w) = k_1 q_1 w q_2 \dots q_m k_2 q_1' w' q_2' \dots q_m' k_3 \dots k_{N+1}$$

and the accepting configuration

$$W = k_1 \bar q_1 \bar q_2 \dots \bar q_m k_2 \bar q_1' \dots k_{N+1}.$$

Here $w', w'', \dots$ are copies of w written in disjoint alphabets.

Let G be the group constructed as in the proof of Theorem 3.2 (see Section 3.3) by imposing the hub relation $W = 1$ on $\mathcal{S}$. Then the word $K(w)$ is equal to 1 in G if and only if $w \in R$.

Consider now another S-machine $\mathcal{S}'$ with input configuration

$$K'(w) = k_1 q_1 q_2 \dots q_m k_2 q_1' w' q_2' \dots q_m' k_3 \dots k_{N+1}.$$

That machine works exactly like $\mathcal{S}$ in the part of the word between k_2 and k_{N+1}, and does nothing in the part between k_1 and k_2. Let G' be the group obtained by imposing the relation $W = 1$ on S'. Then $K'(w) = 1$ in G' if and only if $w \in R$.

Finally consider the amalgamated product $\mathcal{G} = G *_A G'$ where A is generated by all state and tape letters that appear in $K'(W)$. In that group, for every $w \in R$, both $K(w) = 1$ and $K'(w) = 1$. Hence $w = 1$. Thus there exists a natural homomorphism from H into $\mathcal{G}$. It is possible (and not too hard) to prove that this homomorphism is injective. Hence H is inside a finitely presented group G.

Another version of embedding used in [6] employs the so called *Aanderaa trick* [1]: instead of the amalgamated product, we used an HNN extension (see also the survey [45]).

Ones the embedding is established, it is important to understand which properties of a group H can be preserved.

4.2. Dehn functions and quasi-isometric Higman embeddings. First results have been obtained by Clapham [12] and Valiev [53] (see [46] for the history of these results): they proved that the solvability (even recursively enumerable degree) of the word problem and the level in the polynomial hierarchy of the word problem is preserved under some versions of Higman embedding.

In [6], Birget, Olshanskii, Rips and the author of this paper obtained a much stronger result.

Theorem 4.1 ([6]). *Let H be a finitely generated group with word problem solvable by a non-deterministic Turing machine with time function $\leq T(n)$ such that $T(n)^4$ is superadditive. Then H can be embedded into a finitely presented group G with Dehn function $\leq n^2 T(n^2)^4$ in such a way that H has bounded distortion in G.*

This theorem immediately implies the following characterization of groups with word problem in *NP*.

Theorem 4.2 ([6]). *A finitely generated group H has word problem in NP if and only if H is embedded quasi-isometrically into a finitely presented group with polynomial Dehn function.*

Note that the "if" part of this theorem is trivial: if a finitely generated group is a (not necessarily quasi-isometric) subgroup of a group with polynomial Dehn function, its word problem is in *NP*. The converse part is highly non-trivial, although one can prove that the embedding described in Section 4.1 satisfies the desired properties (in [6], we used the Aanderaa trick).

From the logic point of view, Theorem 4.2 means that for every (arbitrary clever) algorithm solving the word problem in a finitely generated group, there exists a finitely presented group $G > H$ such that the word problem in H (and, moreover, in G) can be solved by the Miller machine $M(G)$ in approximately the same time as the initial algorithm.

4.3. Preserving the solvability of the conjugacy problem. The conjugacy problem turned out to be much harder to preserve under embeddings. Collins and Miller [14] and Gorjaga and Kirkinskiĭ[20] proved that even subgroups of index 2 of finitely presented groups do not inherit solvability or unsolvability of the conjugacy problem.

In 1976 D. Collins [31] posed the following question (Problem 5.22): *Does there exist a version of the Higman embedding theorem in which the degree of unsolvability of the conjugacy problem is preserved?* In [46], [47] we solved this problem affirmatively. In particular, we proved the following results.

Theorem 4.3 ([46]). *A finitely generated group H has solvable conjugacy problem if and only if it is Frattini embedded into a finitely presented group G with solvable conjugacy problem.*

Theorem 4.4 ([47]). *Every countable recursively presented group with solvable word and power problems is embeddable into a finitely presented group with solvable conjugacy and power problem.*

Recall that a subgroup H of a group G is Frattini embedded in G if every two elements of H that are conjugate in G are also conjugate inside H. We say that G has solvable *power problem* if there exists an algorithm which, given u, v in G says if $v = u^n$ for some $n \neq 0$.

Theorem 4.4 is a relatively easy application of Theorem 4.3.

The construction in [46] is much more complicated than in [6]. First we embed H into a finitely presented group H_1 preserving the solvability of the word problem. Then we use the Miller S-machine $M(H_1)$ to solve the word problem in H. In order to overcome technical difficulties, we needed certain parts of words appearing the computation to be always positive. The standard positivity checkers do not work because they are S-machines as well, and can insert negative letters! So we used some ideas from the original Boone–Novikov proofs. That required introducing new generators, x-letters (in addition to the a-, q-, and θ-letters in S-machines) and Baumslag–Solitar relations. In addition, to analyze the conjugacy problem in G, we had to consider annular diagrams which are more complicated than van Kampen disc diagrams. Different types of annular diagrams (spirals, roles, etc.) required different treatment.

We do not have any reduction of the complexity of the conjugacy problem in H to the complexity of the conjugacy problem in G. In particular, solving the conjugacy problem in G, in some cases required solving systems of equations in free groups (i.e. the Makanin–Razborov algorithm).

5. Non-amenable finitely presented groups

One of the most important applications of S-machines and Higman embeddings so far was the construction of a finitely presented counterexample to the von Neumann

problem, i.e. a finitely presented non-amenable group without non-Abelian free subgroups [48].

5.1. Short history of the problem. Hausdorff [27] proved in 1914 that one can subdivide the 2-sphere minus a countable set of points into 3 parts A, B, C, such that each of these three parts can be obtained from each of the other two parts by a rotation, and the union of two of these parts can be obtained by rotating the third part. This implied that one cannot define a finitely additive measure on the 2-sphere which is invariant under the group SO(3). In 1924 Banach and Tarski [4] generalized Hausdorff's result by proving, in particular, that in $\mathbb{R}^3$, every two bounded sets A, B with non-empty interiors can be decomposed $A = \bigcup_{i=1}^n A_i$, $B = \bigcup_{i=1}^n B_i$ such that A_i can be rotated to B_i, $i = 1, \ldots, n$ (the so called Banach–Tarski paradox). Von Neumann [54] was first who noticed that the cause of the Banach–Tarski paradox is not the geometry of $\mathbb{R}^3$ but an algebraic property of the group SO(3). He introduced the concept of an amenable group (he called such groups "measurable") as a group G which has a left invariant finitely additive measure μ, $\mu(G) = 1$, noticed that if a group is amenable then any set it acts upon freely also has an invariant measure and proved that a group is not amenable provided it contains a free non-Abelian subgroup. He also showed that groups like $\mathrm{PSL}(2, \mathbb{Z})$, $\mathrm{SL}(2, \mathbb{Z})$ contain free non-Abelian subgroups. So analogs of Banach–Tarski paradox can be found in $\mathbb{R}^2$ and even $\mathbb{R}$ (for a suitable group of "symmetries"). Von Neumann showed that the class of amenable groups contains Abelian groups, finite groups and is closed under taking subgroups, extensions, and infinite unions of increasing sequences of groups. Day [15] and Specht [51] showed that this class is closed under homomorphic images. The class of groups without free non-Abelian subgroups is also closed under these operations and contains Abelian and finite groups.

The problem of existence of non-amenable groups without non-Abelian free subgroups probably goes back to von Neumann and became known as the "von Neumann problem" in the fifties. Probably the first paper where this problem was formulated was the paper by Day [15]. It is also mentioned in the monograph by Greenleaf [21] based on his lectures given in Berkeley in 1967. Tits [52] proved that every non-amenable matrix group over a field of characteristic 0 contains a non-Abelian free subgroup. In particular every semisimple Lie group over a field of characteristic 0 contains such a subgroup.

First counterexamples to the von Neumann problem were constructed by Olshanskii [37]. He proved that the Tarsky monsters, both torsion-free and torsion (see [38]), are not amenable. Later Adian [2] showed that the non-cyclic free Burnside group of odd exponent $n \geq 665$ with at least two generators (that is the group given by the presentation $\langle a_1, \ldots, a_m \mid u^n = 1$, where u runs over all words in the alphabet $\{a_1, \ldots, a_m\}\rangle$) is not amenable.

Both Olshanskii's and Adian's examples are not finitely presented: in the modern terminology these groups are inductive limits of word hyperbolic groups, but they are not hyperbolic themselves. Since many mathematicians are mostly interested

in groups acting "nicely" on manifolds, it is natural to ask if there exists a finitely presented non-amenable group without non-Abelian free subgroups. This question was explicitly formulated, for example, by Grigorchuk in [31] and by Cohen in [13]. This question is one of a series of similar questions about finding finitely presented "monsters", i.e. groups with unusual properties. Probably the most famous problem in that series is the (still open) problem about finding a finitely presented infinite torsion group. Other similar problems ask for finitely presented divisible group (group where every element has roots of every degree), finitely presented Tarski monster, etc. In each case a finitely generated example can be constructed as a limit of hyperbolic groups (see [38]), and there is no hope to construct finitely presented examples as such limits.

One difficulty in constructing a finitely presented non-amenable group without free non-Abelian subgroups is that there are "very few" known finitely presented groups without free non-Abelian subgroups. Most non-trivial examples are solvable or "almost" solvable (see [30]), and so they are amenable. The only previously known example of a finitely presented group without free non-Abelian subgroups for which the problem of amenability is non-trivial, is R. Thompson's group F (for the definition of F look in Section 3.2). The question of whether F is not amenable was formulated by R. Geoghegan in 1979. A considerable amount of work has been done to answer this question but it is still open.

5.2. The result. Together with A. Olshanskii, we proved the following theorem.

Theorem 5.1 ([48]). *For every sufficiently large odd n, there exists a finitely presented group $\mathcal{G}$ which satisfies the following conditions.*

1. *$\mathcal{G}$ is an ascending HNN extension of a finitely generated infinite group of exponent n.*

2. *$\mathcal{G}$ is an extension of a non-locally finite group of exponent n by an infinite cyclic group.*

3. *$\mathcal{G}$ contains a subgroup isomorphic to a free Burnside group of exponent n with 2 generators.*

4. *$\mathcal{G}$ is a non-amenable finitely presented group without free non-cyclic subgroups.*

Notice that parts 1 and 3 of Theorem 5.1 immediately imply part 2. By a theorem of Adian [2], part 3 implies that $\mathcal{G}$ is not amenable. Thus parts 1 and 3 imply part 4.

Note that the first example of a finitely presented group which is a cyclic extension of an infinite torsion group was constructed by Grigorchuk [22]. But the torsion subgroup in Grigorchuk's group does not have a bounded exponent and his group is amenable (it was the first example of a finitely presented amenable but not elementary amenable group).

5.3. The proof. Let us present the main ideas of our construction. We first embed the free Burnside group $B(m, n) = \langle \mathcal{B} \rangle$ of odd exponent $n \gg 1$ with $m > 1$ generators $\{b_1, \dots, b_m\} = \mathcal{B}$ into a finitely presented group $\mathcal{G}' = \langle \mathcal{C} \mid \mathcal{R} \rangle$ where $\mathcal{B} \subset \mathcal{C}$. This is done as in Section 4.1 using an S-machine recognizing all words of the form u^n. The advantage of S-machines is that such an S-machine can be easily and explicitly constructed (see [45]). Then we take a copy $\mathcal{A} = \{a_1, \dots, a_m\}$ of the set $\mathcal{B}$, and a new generator $\boldsymbol{t}$, and consider the group given by generators $\mathcal{C} \cup \mathcal{A}$ and the following three sets of relations:

(1) the set $\mathcal{R}$ of the relations of the finitely presented group $\mathcal{G}'$ containing $B(m, n)$;

(2) (u-relations) $y = u_y$, where u_y, $y \in \mathcal{C}$, is a certain word in $\mathcal{A}$ these words satisfy a very strong small cancellation condition; these relations make $\mathcal{G}'$ (and $B(m, n)$) embedded into a finitely presented group generated by $\mathcal{A}$;

(3) ($\boldsymbol{t}$-relations) $\boldsymbol{t}^{-1} a_i \boldsymbol{t} = b_i$, $i = 1, \dots, m$; these relations make $\langle \mathcal{A} \rangle$ a conjugate of its subgroup of exponent n (of course, the group $\langle \mathcal{A} \rangle$ gets factorized).

The resulting group $\mathcal{G}$ is obviously generated by the set $\mathcal{A} \cup \{\boldsymbol{t}\}$ and is an ascending HNN extension of its subgroup $\langle \mathcal{A} \rangle$ with the stable letter $\boldsymbol{t}$. Every element in $\langle \mathcal{A} \rangle$ is a conjugate of an element of $\langle \mathcal{B} \rangle$, so $\langle \mathcal{A} \rangle$ is an m-generated group of exponent n. This immediately implies that $\mathcal{G}$ is an extension of a group of exponent n (the union of increasing sequence of subgroups $\boldsymbol{t}^s \langle \mathcal{A} \rangle \boldsymbol{t}^{-s}$, $s = 1, 2, \dots$) by a cyclic group.

Hence it remains to prove that $\langle \mathcal{A} \rangle$ contains a copy of the free Burnside group $B(2, n)$.

In order to prove that, we construct a list of defining relations of the subgroup $\langle \mathcal{A} \rangle$. As we have pointed out, the subgroup $\langle \mathcal{A} \cup \mathcal{C} \rangle = \langle \mathcal{A} \rangle$ of $\mathcal{G}$ clearly satisfies all *Burnside relations* of the form $v^n = 1$. Thus we can add all Burnside relations

(4) $v^n = 1$ where v is a word in $\mathcal{A} \cup \mathcal{C}$

to the presentation of group $\mathcal{G}$ without changing the group.

If Burnside relations were the only relations in $\mathcal{G}$ among letters from $\mathcal{B}$, the subgroup of $\mathcal{G}$ generated by $\mathcal{B}$ would be isomorphic to the free Burnside group $B(m, n)$ and that would be the end of the story. Unfortunately there are many more relations in the subgroup $\langle \mathcal{B} \rangle$ of $\mathcal{G}$. Indeed, take any relation $r(y_1, \dots, y_s)$, $y_i \in \mathcal{C}$, of $\mathcal{G}$. Using u-relations (2), we can rewrite it as $r(u_1, \dots, u_s) = 1$ where $u_i \equiv u_{y_i}$. Then using $\boldsymbol{t}$-relations, we can substitute each letter a_j in each u_i by the corresponding letter $b_j \in \mathcal{B}$. This gives us a relation $r' = 1$ which will be called a relation *derived* from the relation $r = 1$, the operator producing derived relations will be called the $\boldsymbol{t}$-*operator*. We can apply the $\boldsymbol{t}$-operator again and again producing the second, third, ..., derivatives $r'' = 1, r''' = 1, \dots$ or $r = 1$. We can add all *derived relations*

(5) $r' = 1, r'' = 1, \dots$ for all relations $r \in \mathcal{R}$

to the presentation of $\mathcal{G}$ without changing $\mathcal{G}$.

Now consider the group H generated by $\mathcal{C}$ subject to the relations (1) from $\mathcal{R}$, the Burnside relations (4) and the derived relations (5). The structure of the relations of H immediately implies that H contains subgroups isomorphic to $B(2, n)$. Thus it is enough to show that the natural map from H to $\mathcal{G}$ is an embedding.

The idea is to consider two auxiliary groups. The group $\mathcal{G}_1$ generated by $\mathcal{A} \cup \mathcal{C}$ subject to the relations (1) from $\mathcal{R}$, u-relations (2), the Burnside relations (4), and the derived relations (5). It is clear that $\mathcal{G}_1$ is generated by $\mathcal{A}$ and is given by relations (1) and (5) where every letter $y \in \mathcal{C}$ is replaced by the corresponding word u_y in the alphabet $\mathcal{A}$ plus all Burnside relations (4) in the alphabet $\mathcal{A}$. Let L be the normal subgroup of the free Burnside group $B(\mathcal{A}, n)$ (freely generated by $\mathcal{A}$) generated as a normal subgroup by all relators (1) from $\mathcal{R}$ and all derived relators (5) where letters from $\mathcal{C}$ are replaced by the corresponding words u_y. Then $\mathcal{G}_1$ is isomorphic to $B(\mathcal{A}, n)/L$.

Consider the subgroup U of $B(\mathcal{A}, n)$ generated (as a subgroup) by $\{u_y \mid y \in \mathcal{C}\}$. The words u_y, $y \in \mathcal{C}$, are chosen in such a way that the subgroup U is a free Burnside group freely generated by u_y, $y \in \mathcal{C}$, and it satisfies the *congruence extension* property, namely every normal subgroup of U is the intersection of a normal subgroup of $B(\mathcal{A}, n)$ with U.

All defining relators of $\mathcal{G}_1$ are inside U. Since U satisfies the congruence extension property, the normal subgroup $\bar{L}$ of U generated by these relators is equal to $L \cap U$. Hence $U/\bar{L}$ is a subgroup of $B(\mathcal{A}, n)/L = \mathcal{G}_1$. But by the choice of U, there exists a (natural) isomorphism between U and the free Burnside group $B(\mathcal{C}, n)$ generated by $\mathcal{C}$, and this isomorphism takes $\bar{L}$ to the normal subgroup generated by relators from $\mathcal{R}$ and the derived relations (5). Therefore $U/\bar{L}$ is isomorphic to H (since, by construction, H is generated by $\mathcal{C}$ subject to the Burnside relations, relations from $\mathcal{R}$ and derived relations)! Hence H is a subgroup of $\mathcal{G}_1$. Let $\mathcal{G}_2$ be the subgroup of H generated by $\mathcal{B}$.

Therefore we have

$$\mathcal{G}_1 \geq H \geq \mathcal{G}_2.$$

Notice that the map $a_i \to b_i$, $i = 1, \ldots, m$, can be extended to a homomorphism $\phi_{1,2} \colon \mathcal{G}_1 \to \mathcal{G}_2$. Indeed, as we mentioned above $\mathcal{G}_1$ is generated by $\mathcal{A}$ subject to Burnside relations, all relators from $\mathcal{R}$ and all derived relators (5) where letters from $\mathcal{C}$ are replaced by the corresponding words u_y. If we apply $\phi_{1,2}$ to these relations, we get Burnside relations and derived relations which hold in $\mathcal{G}_2 \leq H$.

The main technical statement of the paper shows that $\phi_{1,2}$ is an isomorphism, that is for every relation $w(b_1, \ldots, b_m)$ of $\mathcal{G}_2$ the relation $w(a_1, \ldots, a_m)$ holds in $\mathcal{G}_1$. This implies that the HNN extension $\langle \mathcal{G}_1, \boldsymbol{t} \mid \boldsymbol{t}^{-1}\mathcal{G}_1\boldsymbol{t} = \mathcal{G}_2 \rangle$ is isomorphic to $\mathcal{G}$. Indeed, this HNN extension is generated by $\mathcal{G}_1$ and $\boldsymbol{t}$, subject to relations (1), (2), (4), (5) of $\mathcal{G}_1$ plus relations (3). So this HNN extension is presented by relations (1)–(5) which is the presentation of $\mathcal{G}$. Therefore $\mathcal{G}_1$ is a subgroup of $\mathcal{G}$, hence H is a subgroup of $\mathcal{G}$ as well.

The proof of the fact that $\phi_{1,2}$ is an isomorphism requires a detailed analysis of the group H. This group can be considered as a factor-group of the group H' generated by $\mathcal{C}$ subject to the relations (1) from $\mathcal{R}$ and derived relations (5) over the normal subgroup generated by Burnside relations (4). In other words, H is the *Burnside factor* of H'.

Burnside factors of free groups have been studied extensively starting with the celebrated paper by Adian and Novikov [36]. Later Olshanskii developed a geometric method of studying these factors in [38]. These methods were extended to arbitrary hyperbolic groups in [39]

The main problem we face in this paper is that H' is "very" non-hyperbolic. In particular, the set of relations $\mathcal{R}$ contains many commutativity relations, so H' contains non-cyclic torsion-free Abelian subgroups which cannot happen in a hyperbolic group.

We use a weak form of relative hyperbolicity that does hold in H'. In order to roughly explain this form of relative hyperbolicity used in the proof, consider the following example. Let $P = F_A \times F_B$ be the direct product of two free groups of rank m. Then the Burnside factor of P is simply $B(m,n) \times B(m,n)$. Nevertheless the theory of [38] cannot be formally applied to P. Indeed, there are arbitrarily thick rectangles corresponding to relations $u^{-1}v^{-1}uv = 1$ in the Cayley graph of P so diagrams over P are not A-maps in the terminology of [38] (i.e. they do not look like hyperbolic spaces). But one can obtain the Burnside factor of P in two steps. First we factorize F_A to obtain $Q = B(m,n) \times F_B$. Since F_A is free, we can simply use [38] to study this factor.

Now we consider all edges labeled by letters from A in the Cayley graph of Q as 0-edges, i.e. edges of length 0. As a result the Cayley graph of Q becomes a hyperbolic space (a tree). This allows us to apply the theory of A-maps from [38] to obtain the Burnside factor of Q. In fact Q is weakly relatively hyperbolic in the sense of our paper [48], i.e. it satisfies conditions (Z1), (Z2), (Z3) from the paper. The class of groups satisfying these conditions is very large and includes groups corresponding to S-machines considered in [48].

Recall that set $\mathcal{C}$ consists of tape letters, state letters, and command letters. In different stages of the proof some of these letters become 0-letters.

Trapezia corresponding to computations of the S-machine play central role in our study of the Burnside factor H of H'. As in [38], the main idea is to construct a graded presentation $\mathcal{R}'$ of the Burnside factor H of H' where longer relations have higher ranks and such that every van Kampen diagram over the presentation of H' has the so called property A from [38]. In all diagrams over the graded presentation of H, cells corresponding to the relations from $\mathcal{R}$ and derived relations are considered as 0-cells or cells of rank $1/2$, and cells corresponding to Burnside relations from the graded presentation are considered as cells of ranks 1, 2,.... So in these van Kampen diagrams "big" Burnside cells are surrounded by "invisible" 0-cells and "small" cells.

The main part of property A from [38] is the property that if a diagram over $\mathcal{R}'$ contains two Burnside cells Π_1, Π_2 connected by a rectangular *contiguity* subdia-

gram Γ of rank 0 where the sides contained in the contours of the two Burnside cells are "long enough" then these two cells cancel, that is the union of Γ, Π, Π' can be replaced by a smaller subdiagram. This is a "graded substitute" to the classic property of small cancellation diagrams (where contiguity subdiagrams contain no cells).

In our case, contiguity subdiagrams of rank 0 turn out to be trapezia (after we clean them of Burnside 0-cells), so properties of contiguity subdiagrams can be translated into properties of the machine $\mathcal{S}$.

References

[1] Aanderaa, S., A proof of Higman's embedding theorem using Britton extensions of groups. In *Word Problems, Decision Problems and the Burnside problem in group theory*, North-Holland, Amsterdam, London 1973, 1–18.

[2] Adian, S. I., Random walks on free periodic groups. *Izv. Akad. Nauk SSSR Ser. Mat.* **46** (1982), 1139–1149.

[3] Allcock, D., An isoperimetric inequality for the Heisenberg groups. *Geom. Funct. Anal.* **8** (2) (1998), 219–233.

[4] Banach, S., Tarski, A., Sur la décomposition de ensembles de points an parties respectivement congruentes. *Fund. Math.* **6** (1924), 244–277.

[5] Baumslag, G., Miller III, C. F., Short, H., Isoperimetric inequalities and the homology of groups. *Invent. Math.* **113** (1993), 531–560.

[6] Birget, J.-C., Olshanskii, A. Y., Rips, E., Sapir, M. V., Isoperimetric functions of groups and computational complexity of the word problem. *Ann. of Math.* (2) **156** (2002), 467–518.

[7] Bogopolski, O., Martino, A., Maslakova, O., Ventura, E., Free-by-cyclic groups have solvable conjugacy problem. arXive, math.GR/0405178, 2004.

[8] Boone, W. W., Certain simple unsolvable problems in group theory I–VI. *Nederl. Akad. Wetensch. Proc. Ser. A* **57** (1954), 231–237; **57** (1954), 492–497; **58** (1955), 252–256; **58** (1955), 571–577; **60** (1957), 22–27; **60** (1957), 222–232.

[9] Brady, N., Bridson, M. R., There is only one gap in the isoperimetric spectrum. *Geom. Funct. Anal.* **10** (5) (2000), 1053–1070.

[10] Bridson, M. R., Fractional isoperimetric inequalities and subgroup distortion. *J. Amer. Math. Soc.* **12** (4) (1999), 1103–1118.

[11] Bridson, M. R., Groves, D., Free-group automorphisms, train tracks and the beaded decomposition. arXiv math.GR/0507589, 2005.

[12] Clapham, C. R. J., An embedding theorem for finitely generated groups. *Proc. London. Math. Soc.* (3) **17** (1967), 419–430.

[13] Cohen, J. M., Cogrowth and amenability of discrete groups. *J. Funct. Anal.* **48** (3) (1982), 301–309.

[14] Collins, D. J., Miller III, C. F., The conjugacy problem and subgroups of finite index. *Proc. London Math. Soc.* (3) **34** (3) (1977), 535–556.

[15] Day, Mahlon M., Amenable semigroups. *Illinois J. Math.* **1** (1957), 509–544.

[16] Druţu, C., Quasi-isometry invariants and asymptotic cones. International Conference on Geometric and Combinatorial Methods in Group Theory and Semigroup Theory (Lincoln, NE, 2000), *Internat. J. Algebra Comput.* **12** (1–2) (2002), 99–135.

[17] Druţu, C., Filling in solvable groups and in lattices in semisimple groups. *Topology* **43** (5) (2004), 983–1033.

[18] Druţu, C., Sapir, M. V., Tree-graded spaces and asymptotic cones of groups. With an appendix by Denis Osin and Mark Sapir. *Topology* **44** (5) (2005), 959–1058.

[19] Gersten, S. M., Dehn functions and l_1-norms for finite presentations. In *Algorithms and Classification in Combinatorial Group Theory* (ed. by G. Baumslag, C. F. Miller), Math. Sci. Res. Inst. Publ. 23, Springer-Verlag, New York 1992.

[20] Gorjaga, A. V., Kirkinskiĭ, A. S., The decidability of the conjugacy problem cannot be transferred to finite extensions of groups. *Algebra i Logika* **14** (4) (1975), 393–406 (in Russian).

[21] Greenleaf, F. P., *Invariant means on topological groups and their applications.* Van Nostrand Reinhold, New York 1969.

[22] Grigorchuk, R. I., An example of a finitely presented amenable group that does not belong to the class EG. *Mat. Sb.* **189** (1) (1998), 79–100.

[23] Gromov, M., Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[24] Gromov, M., Asymptotic invariants of infinite groups. In *Geometric Group Theory* (ed. by G. A. Niblo and M. A. Roller), Volume 2, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge 1993, 1–295.

[25] Guba, V. S., The Dehn Function of Richard Thompson's Group F is Quadratic. *Invent. Math.* **163** (2006), 313–342.

[26] Guba, V. S., Sapir, M. V., On Dehn functions of free products of groups. *Proc. Amer. Math. Soc.* **127** (7) (1999), 1885–1891.

[27] Hausdorff, F., *Grundzüge der Mengenlehre*. Veit & Company, Leipzig 1914.

[28] Higman, G., Subgroups of finitely presented groups. *Proc. Roy. Soc. Ser. A* **262** (1961), 455–475.

[29] Kapovich, M., Kleiner, B., Geometry of quasi-planes. Preprint, 2004.

[30] Kharlampovich, O. G., Sapir, M. V., Algorithmic problems in varieties. *Internat. J. Algebra Comput.* **5** (4–5) (1995), 379–602.

[31] *Kourovka Notebook. Unsolved Problems in Group Theory*. 5th edition, Novosibirsk 1976.

[32] Kramer, L., Shelah, S., Tent, K., Thomas, S., Asymptotic cones of finitely presented groups. *Adv. Math.* **193** (1) (2005), 142–173.

[33] Madlener, K., Otto, F., Pseudo-natural algorithms for the word problem for finitely presented monoids and groups. *J. Symbolic Comput.* **1** (1989), 383–418.

[34] Miller III, C. F., *On group-theoretic decision problems and their classification*. Ann. of Math. Stud. 68, Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo 1971.

[35] Novikov, P. S., *On the algorithmic insolvability of the word problem in group theory*. Trudy Mat. Inst. im. Steklov. no. 44, Izdat. Akad. Nauk SSSR, Moscow 1955, 143 pp; English transl. Amer. Math. Soc. Transl. Ser. 9, Amer. Math. Soc., Providence, R. I., 1958, pp. 1–122.

[36] Novikov, P. S., Adian, S. I., Infinite periodic groups. I, II, III. *Izv. Akad. Nauk SSSR Ser. Mat.* **32** (1968) 212–244, 251–524, 709–731 (in Russian).

[37] Olshanskii, A. Yu., On the question of the existence of an invariant mean on a group. *Uspekhi Mat. Nauk* **35** (4) (1980), 199–200; English transl. *Russian Math. Surveys* **35** (4) (1980), 180–181.

[38] Olshanskii, A. Yu., *The geometry of defining relations in groups*. Nauka, Moscow 1989.

[39] Olshanskii, A. Yu., The SQ-universality of hyperbolic groups. *Mat. Sb.* **186** (8) (1995), 119–132; English transl. *Sb. Math.* **186** (8) (1995), 1199–1211.

[40] Olshanskii, A. Yu., Groups with quadratic-non-quadratic Dehn functions. arXiv, math.GR/0504349, 2005.

[41] Olshanskii, A. Yu., Sapir, M. V., A finitely presented group with two non-homeomorphic asymptotic cones. arXiv math.GR/0504350, 2005.

[42] Olshanskii, A. Yu., Sapir, M. V., Groups with small Dehn functions and bipartite chord diagrams. *Geom. Funct. Anal.*, to appear; arXiv, math.GR 0411174, 2004.

[43] Olshanskii, A. Yu., Sapir, M. V., Quadratic isometric functions of the Heisenberg groups. A combinatorial proof. *J. Math. Sci.* (*New York*) **93** (6) (1999), 921–927.

[44] Olshanskii, A. Yu., Sapir, M. V., Groups with non-simply connected asymptotic cones. arXiv, math.GR/0501542, 2005.

[45] Olshanskii, A. Yu., Sapir, M. V., Length and area functions on groups and quasi-isometric Higman embeddings. *Internat. J. Algebra Comput.* **11** (2) (2001), 137–170.

[46] Olshanskii, A. Yu., Sapir, M. V., The conjugacy problem and Higman embeddings. *Mem. Amer. Math. Soc.* **170** (804) (2004).

[47] Olshanskii, A. Yu., Sapir, M. V., Subgroups of finitely presented groups with solvable conjugacy problem. arXiv, math.GR/0405337, 2004.

[48] Olshanskii, A. Yu., Sapir, M. V., Non-amenable finitely presented torsion-by-cyclic groups. *Inst. Hautes Études Sci. Publ. Math.* **96** (2002), 43–169.

[49] Papasoglu, P., On the asymptotic cone of groups satisfying a quadratic isoperimetric inequality. *J. Differential Geom.* **44** (1996), 789–806.

[50] Sapir, M. V., Birget, J. C., Rips, E., Isoperimetric and isodiametric functions of groups. *Ann. of Math.* (2) **157** (2002), 345–466.

[51] Specht, W., Zur Theorie der messbaren Gruppen. *Math. Z.* **74** (1960), 325–366.

[52] Tits, J., Free subgroups of linear groups. *J. Algebra* **20** (1972), 250–270.

[53] Valiev, M. K., On polynomial reducibility of the word problem under embedding of recursively presented groups in finitely presented groups. In *Mathematical Foundations of Computer Science* 1975 (ed. by J. Bečvář), Lecture Notes in Comput. Sci. 32, Springer-Verlag, Berlin 1975, 432–438.

[54] von Neumann, J., Zur allgemeinen Theorie des Maßes. *Fund. Math.* **13** (1929), 73–116.

Department of Mathematics, Vanderbilt University, Nashville, TN 37240, U.S.A.
E-mail: m.sapir@vanderbilt.edu

A unified approach to computations with permutation and matrix groups

Ákos Seress*

Abstract. We survey algorithms to compute with large finite permutation and matrix groups. Particular attention will be given to handling both types of groups with similar methods, using structural properties to answer even basic questions such as the order of the input group.

Mathematics Subject Classification (2000). Primary 20B40; Secondary 20B15.

Keywords. Computational group theory, permutation group algorithm, matrix group algorithm.

1. Introduction

There are two basic methods to input a group into a computer: (a) by a presentation, using abstract generators and relations, and (b) by a "concrete" representation as a permutation group or matrix group, defined by a set of generating permutations or matrices. In this survey, we are concerned with groups given as in (b), and with black-box groups (see Definition 3.1), which are a common generalization of permutation and matrix groups. We shall concentrate on the basic questions how to determine the order of the input group G, how to set up a data structure to test membership in G, and how to compute a composition series. For readers interested in a broader range of topics, we recommend our brief survey [47] describing all areas of computational group theory and the more thorough coverage in the recent book [29]. Two monographs providing a comprehensive coverage of the subareas finitely presented groups and permutation groups are [51] and [48], respectively.

Given a set X of permutations or of invertible matrices over a finite field, X generates a finite group G so the undecidability issues related to generator-relator presentations and to infinite matrix groups do not arise. However, $|G|$ can be exponentially large in terms of the input length, so brute-force methods like listing all elements of G are out of question and we have to design efficient algorithms to deal with G.

There are two widely accepted notions of efficiency. On one hand, in practice, it means that we obtain results in reasonable time in the actual computations we perform. On the other hand, in theory, efficiency means fast asymptotic running time. Historically, group computations developed on these two tracks separately,

*Partially supported by the NSA and the NSF.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

but in the last fifteen years or so the two approaches have started to converge. This convergence is not surprising. As we deal with larger and larger inputs, only those practical methods that are asymptotically efficient survive; conversely, the rigorous complexity analysis inspires new algorithms that may have practical implementations. This unification of theory and practice is one of the aspects the title of this paper refers to. Implementations of asymptotically fast algorithms are finding their way into *GAP* [28] and MAGMA [19], the two large computer algebra systems for group computations.

There are two other aspects of unification. One of them is within the matrix group setting. As we shall discuss in Section 3, there are two approaches to matrix group computations and we shall mention the recent efforts to combine them. The other aspect is the uniform treatment of permutation groups and matrix groups, by breaking them into manageable pieces as the image and kernel of appropriate group homomorphisms. This approach is the standard one for matrix groups, but a recently developed data structure enables us to handle both types of groups the same way.

The most recent ICM talk about computational group theory was given eight years ago, by Bill Kantor [30]. His major emphasis was the use of the classification of finite simple groups (CFSG) in the topic and the handling of simple groups. We shall also report the latest developments in simple group management, but this paper is more focused on the *reduction* to the simple group case. There are interesting and deep mathematical problems in both subareas, although the management of simple groups requires mostly group theoretical arguments while the reduction to the simple group case uses a mixture of group theoretic, combinatorial and computer science (design of data structures) methods. Consequences of CFSG are required in the analysis of many reduction algorithms as well.

2. Permutation groups

The fundamental data structures for computation with permutation groups were introduced by Sims [50]; they are called base and strong generating set (SGS). A *base* of $G \leq \mathrm{Sym}(\Omega)$ is a sequence of points $B = (\beta_1, \ldots, \beta_m)$ from Ω such that the pointwise stabiliser $G_B = 1$. A base B naturally defines a subgroup chain

$$G = G^{[1]} \geq G^{[2]} \geq \cdots \geq G^{[m]} \geq G^{[m+1]} = 1$$

where $G^{[i]} := G_{(\beta_1,\ldots,\beta_{i-1})}$ is the pointwise stabilizer of $\{\beta_1, \ldots, \beta_{i-1}\}$. The base is called *non-redundant* if $G^{[i+1]}$ is a proper subgroup of $G^{[i]}$ for all $i \leq m$. For a non-redundant base B, we have $\log |G|/\log N \leq |B| \leq \log |G|$, where $|\Omega| = N$. (As usual in complexity theory, we write logarithms to base 2.)

A *strong generating set* (SGS) for G relative to B is a generating set S for G with the property that

$$\langle S \cap G^{[i]} \rangle = G^{[i]}, \quad \text{for } 1 \leq i \leq m+1.$$

Given $G = \langle X \rangle \leq \mathrm{Sym}(\Omega)$, Sims's algorithm constructs a non-redundant base B and an SGS S relative to B. Once S is known, it is easy to construct (right) transversals T_i

for $G^{[i]}$ mod $G^{[i+1]}$. Crucially, $|T_i| \leq N$ is "small". For all γ in the orbit $\beta_i^{G^{[i]}}$, the transversal T_i contains some $r_\gamma \in G^{[i]}$ with $\beta_i^{r_\gamma} = \gamma$. These transversals can be used to compute $|G| = \prod_{i=1}^{m} |T_i|$ and to factor any $g \in G$ as a product $g = r_m \dots r_1$, for some $r_i \in T_i$. This factorization is unique and it can be done by an efficient algorithm called *sifting*. First, we take $r_1 \in T_1$ such that $\beta_1^g = \beta_1^{r_1}$. Then $g_2 := gr_1^{-1} \in G^{[2]}$, and we can take $r_2 \in T_2$ such that $\beta_2^{g_2} = \beta_2^{r_2}$, etc. Sifting can be used to test membership in G. For details, we refer to [48, Ch. 4].

Sims's algorithm is based on elementary group theory. The running time of the asymptotically fastest versions is $O(|X|N^2 \log^c |G|)$ for some absolute constant c. A factor N can be shaved off the running time by randomization:

Theorem 2.1 ([7]). *Given $G = \langle X \rangle \leq \mathrm{Sym}(\Omega)$ with $|\Omega| = N$, a base B and an SGS relative to B can be computed in $O(|X|N \log^c |G|)$ time, by a Monte Carlo algorithm with an arbitrarily small positive but fixed upper bound on the probability of incorrect output.*

Recall that a randomized algorithm is called *Monte Carlo* if there is a chance of an incorrect output but an upper bound for the probability of error can be prescribed by the user. On the contrary, a *Las Vegas* algorithm never returns an incorrect answer but it may report failure with probability bounded by the user.

The algorithm in Theorem 2.1 is still elementary. The quadratic $O(N^2)$ bottlenecks are broken by randomization, and by combinatorial tricks like working with base images instead of full permutations or to test membership in certain large subsets of G without listing those subsets. If $\log |G|$ is bounded from above by a polylogarithmic, $\log^c N$, function of N then the running time is a *nearly linear*, $O((|X|N) \log^{c'} (|X|N))$, function of the input length $|X|N$. This motivated the following definition. An infinite family $\mathcal{G}$ of permutation groups is called *small-base* if every group $G \in \mathcal{G}$ of degree m satisfies $\log |G| < \log^c m$ for some fixed constant c. Important families of groups are small-base, including all permutation representations of non-alternating simple groups.

The algorithm in Theorem 2.1 is also practical. In *GAP*, currently permutation group computations are based on an implementation of this algorithm. The certainty of a correct answer, if desired, is obtained by a quadratic algorithm of Sims that checks the correctness of a base and SGS (see [48, Section 8.2]).

For arbitrary inputs, where $\log |G|$ may become comparable to N, no deterministic version of Sims's algorithm is known to run faster than $O(N^5 + |X|N^2)$. Fortunately, ideas from a purely theoretical development come to the rescue. In [9], an algorithm is described to handle permutation groups in the parallel computational model NC. Informally, in NC we can work with polynomially many, n^c, processors, but we have only polylogarithmic, $\log^c n$, time in terms of the input length n. Note that sifting is inherently sequential (we have to know the coset representatives $r_1, \dots, r_i$ before r_{i+1} can be computed in a factorization $g = r_m \dots r_1$) so a Sims-based approach may work in NC only for small-base groups. The algorithm in [9] is based on entirely different

principles, exploring the structure of the input group G. By the time it computes $|G|$, it also obtains a composition series for G. Some of these ideas were also used in the more realistic domain of sequential computations to break the long-standing $O(N^5)$ barrier:

Theorem 2.2 ([10]). *Given $G = \langle X \rangle \leq \mathrm{Sym}(\Omega)$ with $|\Omega| = N$, there is a deterministic algorithm with $O(N^4 \log^c N + |X|N^2)$ running time to compute $|G|$ and to set up a data structure for testing membership in G.*

The algorithm in Theorem 2.2 detects all large alternating composition factors of G and handles them by special methods, while the rest of the group is handled using Sims's ideas.

How can we detect large alternating sections in a permutation group? Combinatorial reduction (action on orbits, and then action on blocks of imprimitivity) leads to primitive permutation groups. At that point, we invoke a consequence of CFSG [20]: any primitive permutation group $H \leq \mathrm{Sym}(\Delta)$ of degree n is a small-base group, unless $n = \binom{m}{k}^r$ for some positive integers m, k, r and Δ can be identified with r-tuples of k-sets of an m-element set, $A_m^r \leq H \leq S_m \wr S_r$, and H acts naturally on these sequences in the so-called product action of wreath products. Such an H is called a *group of Cameron type* in [10]. In this very special situation, [10] gives a combinatorial algorithm to construct a collection Σ of mr subsets of Δ so that H acts on Σ in the natural imprimitive action of wreath products.

The input group G is handled by a recursive procedure. We define a homomorphism $\varphi\colon G \to \mathrm{Sym}(\Delta)$ for some Δ, process $\mathrm{Im}(\varphi)$, obtain generators for $\mathrm{Ker}(\varphi)$, and process $\mathrm{Ker}(\varphi)$. If G is transitive then Δ is an orbit and $\mathrm{Im}(\varphi)$ is the restriction of G to this orbit; and if G is transitive but imprimitive then Δ is a block system and $\mathrm{Im}(\varphi)$ is the action on this block system. When the recursion arrives to a primitive group then we test whether it is of Cameron type. If not, then Sims's base-SGS method is used to process it. If it is of Cameron type then a further homomorphism is constructed to the natural imprimitive action. The full alternating and symmetric groups A_m and S_m, encountered in their natural action on m points, are handled by combinatorial methods, as a special case of the constructive recognition of almost simple groups (see Section 3.1).

We finish this section by announcing a Las Vegas upgrade of Theorem 2.1.

Theorem 2.3 ([32], [33]). *Given $G = \langle X \rangle \leq \mathrm{Sym}(\Omega)$ with $|\Omega| = N$ and G having no composition factors of Lie type ${}^2G_2$ and ${}^2F_4$, a base and SGS for G can be computed in $O(|X|N \log^c |G|)$ time, by a Las Vegas algorithm with an arbitrarily small but fixed positive upper bound on the probability of incorrect output.*

Although the statements of Theorems 2.1 and 2.3 are similar, the proofs are based on entirely different principles. The algorithm in Theorem 2.3 computes a composition series for G by a Monte Carlo algorithm, recognizes constructively the composition factors (see Section 3.1), and then uses the isomorphisms set up by the recognition algorithms to write a presentation for G. Evaluation of this presentation verifies

the correctness of the entire computation. The groups ${}^2G_2(q)$ have to be excluded because currently it is not known that they have short presentations suitable for the time requirement of this application; the groups ${}^2F_4(q)$ are excluded because the known constructive recognition algorithm is not fast enough.

There is a large library of Monte Carlo algorithms that run in nearly linear time for small-base inputs (see [48, Ch. 5 and 6]). The significance of Theorem 2.3 is that if the initial base-SGS computation is correct for some input group G then all of these nearly linear-time algorithms are *automatically upgraded to Las Vegas*.

Summarizing, we saw that the basic tasks of finding the order and setting up membership testing in permutation groups can be performed by elementary methods in polynomial time, but randomization and the structural exploration of the input group provide much faster algorithms.

3. Matrix groups

The basic problems for matrix groups over finite fields, such as membership and order, seem to be much harder than the corresponding problems for permutation groups. The fundamental difference is that there is no longer, in general, a decreasing sequence of subgroups from G to 1 in which all successive indices are small; this makes an analogue of Sims' base-SGS approach infeasible. For permutation groups, the natural divide-and-conquer approach leads to primitive groups, and those groups can be reduced to symmetric groups or else they are small-base groups. In contrast, a large variety of primitive irreducible matrix groups has order $\exp(\Omega(d^2))$ (here d is the dimension of the matrices). Finally, even for 1×1 matrices, the problems are closely related to discrete logarithm computations.

For later use, we define two versions of the discrete logarithm problem:

(DL1) Given $a, b \in \mathrm{GF}(q)^*$, determine whether $a \in \langle b \rangle$.

(DL2) Given $a, b \in \mathrm{GF}(q)^*$, determine whether $a \in \langle b \rangle$. If the answer is yes then find an exponent x such that $a = b^x$.

Finding the order of $G = \langle X \rangle \leq \mathrm{GL}(1, q)$ is between these two problems in difficulty: version (DL1) can be reduced to it in polynomial time, while it can be reduced to version (DL2) in polynomial time. We note that neither version of the discrete logarithm problem has at present a polynomial-time solution, although subexponential algorithms exist even for the more difficult version (DL2) [39].

Despite all the difficulties listed above, significant progress has been made recently on matrix groups and associated data structures, and currently this is the most active area of computational group theory. However, contrary to the permutation group case, it seems that randomization and a full structural exploration of the input is not only a speedup, but an essential and unavoidable tool. This means that we have to set up a

recursive scheme of homomorphisms, breaking the input into the image and kernel. This reduction bottoms out at matrix groups H that are almost simple modulo scalars. At these terminal stages of the recursion, we have to find the name of the isomorphism type of H, and then set up an identification with a standard permutation or matrix representation of this isomorphism type.

First, we discuss the methods for handling almost simple groups. For that, we need two definitions.

Definition 3.1. A *black-box group* is a group whose elements are encoded as words of length at most N over some alphabet T and some bound N. Not every word represents a group element and the same group element may be represented by more than one word. Moreover, an oracle (the "black box") performs the following three operations: given (words representing) $g, h \in G$, it can compute (a word representing) gh, g^{-1}, and it can decide whether $g = 1$.

Our definition is slightly more general than the original one [11], where only 0-1 strings of uniform lengths are allowed. The primary examples of black-box groups are permutation groups and matrix groups, but there are two other important examples. One of them is a power-conjugate presentation for a finite solvable group, where each group element has a canonical form $a_1^{e_1} a_2^{e_2} \dots a_m^{e_m}$ for a suitable generating sequence $(a_1, \dots, a_m)$. More important for our present discussion is that permutation groups G can be considered as black-box groups where the alphabet is an SGS for G (see [48, Section 5.3]). In small-base groups, group operations using words in the strong generators are asymptotically much faster than permutation multiplications. These special types of black-box groups play an important role, for example, in the proof of Theorem 2.3. In these black-box groups, we lose all information stored implicitly in the cycle structure of permutations, and algorithms can utilize only the three black-box operations defined above.

In some situations, we also consider permutation and matrix groups with their natural group operations as black-box groups, because permutation group theoretic notions like orbits or cycle structure, or geometric notions like invariant subspaces or characteristic polynomials in the matrix group case, do not help. For example, we have no better methods for generating random elements in a matrix group than creating new group elements from the given generators by multiplications and inversions, that is, by black-box operations [4], [22]. Black-box group algorithms, with the natural permutation operations, are also used in computations of normal closures, derived series, and related algorithms both in theory [23], [6] and in *GAP*.

The second definition we require is of a straight-line program (SLP). It is a data structure to circumvent problems with overly long words in generators.

Definition 3.2. Given $G = \langle X \rangle$, a *straight-line program of length m* reaching some $g \in G$ is a sequence of expressions $(w_1, \dots, w_m)$ such that for each i one of the following holds: w_i is a symbol for some element of X, $w_i = (w_j, -1)$ for some $j < i$, or $w_i = (w_j, w_k)$ for some $j, k < i$, such that, if the expressions are evaluated,

then the value of w_m is g. Here, $(w_j, -1)$ is evaluated as the inverse of the evaluated value of w_j, and (w_j, w_k) is evaluated as the product of the evaluated values of w_j and w_k.

3.1. Recognition of almost simple groups. Now we are ready to discuss matrix groups that are almost simple modulo scalars or, more generally, almost simple black-box groups. There are two basic tasks: non-constructive recognition and constructive recognition. *Non-constructive recognition* of an almost simple group means to name the isomorphism type. The first such algorithm is in [41], where it is decided whether a given group $G \leq \mathrm{GL}(d, p^e)$ contains $\mathrm{SL}(d, p^e)$. A sample of random elements is taken, and we look for elements whose order is divisible by some primitive prime divisor (ppd) of $p^{de} - 1$ and of $p^{(d-1)e} - 1$. (Recall that a *primitive prime divisor* of $p^n - 1$ is a prime $r \mid p^n - 1$ which does not divide $p^i - 1$ for any $i < n$.) If $G \geq \mathrm{SL}(d, p^e)$ then elementary estimates show that both kinds of ppd's occur frequently enough so that a small sample of random elements detects them; however, CFSG is invoked to prove that if both kinds of ppd's occur then indeed $G \geq \mathrm{SL}(d, p^e)$. Subsequently, similar algorithms were designed to recognize the other classical groups in their natural matrix representations [21], [43]. The culmination of this type of results is in [8]: given an almost simple black-box group G of Lie type and given the characteristic p of G, the isomorphism type of G can be computed. This algorithms still looks for elements whose order is divisible by various ppd's and pairs of ppd's. We note that the ppd property can be checked in the black-box group setting, without computing element orders. If we know only that a matrix group is simple of Lie type modulo scalar matrices, its characteristic can be determined by recent algorithms in [38] and [49]. All algorithms mentioned in this paragraph are Monte Carlo with polynomial running time, and have efficient implementations.

In applications in recursive schemes breaking down arbitrary matrix groups to almost simple pieces, non-constructive recognition is not sufficient; we need the more involved constructive recognition. Given an almost simple black-box group $G = \langle X \rangle$, *constructive recognition* of G is a Las Vegas algorithm that, besides naming the isomorphism type of G, computes an isomorphism $\varphi\colon G \to C$ with a standard permutation representation or (projective) matrix representation of this isomorphism type. The isomorphism φ is defined by giving the images of a new generating set $Y \leq G$. Moreover, we require that, given any $g \in G$, a short SLP reaching g from Y can be computed, and given any $h \in C$, $\varphi^{-1}(h)$ can be computed.

The first constructive recognition algorithm, for black-box groups $\mathrm{GL}(n, 2)$, was given in [24]. Subsequently, constructive recognition of all classical groups (in [32]) and exceptional groups G (in [31]) was accomplished. These algorithms require the characteristic p of G as part of the input. The rough idea is the following. Since in the black-box setting we do not have a vector space to work with, the algorithms construct a large elementary abelian p-section P of G such that the conjugation action of a maximal parabolic $H \leq G$ on P is isomorphic to the natural matrix action of H, and subsequently extend this matrix action to arbitrary elements of G. Constructive

recognition of alternating and symmetric groups is much easier [12], [13], [14]. A recent algorithm [2] recognizes constructively about half of the sporadic groups, using a generalization of Sims's sifting through subset chains instead of subgroup chains.

An exciting new method by Ryba toward the constructive recognition of some Lie-type groups is described in [45] and [46]. Let p be an odd prime, and let G be a group of untwisted Lie type defined over a field of characteristic p. Suppose further that the associated Lie algebra of G is simple. Then, given any absolutely irreducible characteristic p representation $G = \langle X \rangle$, a polynomial-time Las Vegas algorithm computes the action of the generator set X on the Chevalley basis of the Lie algebra of G. Hence the constructive recognition problem is reduced to consideration of the adjoint representation.

Ryba is currently working on the extension of this algorithm to all Lie-type groups. Although his methods are still under development, they seem to have the potential to become the major tool of constructive recognition, reducing the use of the methodology of [32], [31] only to the natural and cross-characteristic representations of Lie-type groups.

The running times of the algorithms in [32], [31] are polynomials in the rank r and the defining field size q of the Lie-type input group G. However, the length of the input may be only $O(r^2 \log q)$, so for large q the running time is exponential. An idea to overcome this difficulty is in [25], where the groups $\mathrm{SL}(2, q)$ in their standard 2×2 matrix setting are recognized in polynomial time of the input length *plus* polynomially many calls to an oracle solving version (DL2) of the discrete logarithm problem in $\mathrm{GF}(q)$. Later this algorithm was extended to arbitrary matrix representations of $\mathrm{PSL}(2, q)$ [26]. This motivated the following definition.

Definition 3.3. Let G be an almost simple group of Lie type defined over the field $\mathrm{GF}(p^l)$. We say that G is *constructively recognizable with a discrete logarithm oracle*, in short G is *CRDLO*, if for *any* quasisimple representation of G in characteristic p, G can be constructively recognized in time polynomial in the input length plus the time of polynomially many calls to a discrete logarithm oracle in $\mathrm{GF}(p^l)$. (Note that the field of definition $\mathrm{GF}(p^l)$ may be different from the field $\mathrm{GF}(p^e)$ over which the input matrices are given. Recall that a group G is called *quasisimple* if $G/Z(G)$ is simple and G equals its derived subgroup.)

Theorem 3.4 ([16], [15], [17]). *All classical groups are CRDLO.*

The algorithms of this theorem are based on [32], using an oracle to handle $\mathrm{SL}(2, q)$ subgroups. In turn, this oracle is based on the methods of [26]. The case of special linear, symplectic, and unitary groups is a more or less straightforward modification of [32], but the case of orthogonal groups involves significant additional technical difficulties.

3.2. The general case. Now we turn to the case of arbitrary matrix groups. There are two basic methods for the breakup of the input into manageable pieces (which,

in most cases, amounts to the reduction to almost simple groups). The *geometric approach*, summarized in [37], is based on Aschbacher's classification of matrix groups [3]. This classification defines eight types of geometric subgroups of $\mathrm{GL}(d, q)$, and groups $G \leq \mathrm{GL}(d, q)$ belonging to seven of these types have a naturally associated $N \lhd G$ which enables the recursive handling of G/N and N. These classes consist of reducible groups, imprimitive groups, normalizers of extraspecial groups, and so on. For example, in the case of reducible groups we can consider the homomorphism defined by the restriction to the action on an invariant subspace, and the kernel of this action. The eighth geometric category contains the classical groups in their natural representation. The groups not belonging to any of the eight geometric categories are almost simple modulo scalars. After contributions by many people (see [44] for an overview), O'Brien has a working implementation of the reduction to the almost simple case.

By contrast, the *black-box group approach*, initiated by Babai and Beals [12], tries to determine the abstract group-theoretic structure of G. Every finite group G has a series of characteristic subgroups $1 \leq M_1 \leq M_2 \leq M_3 \leq G$, where M_1 is solvable, M_2/M_1 is isomorphic to a direct product $T_1 \times \cdots \times T_k$ of nonabelian simple groups, M_3/M_2 is solvable, and G/M_3 is a permutation group, permuting the simple groups T_i. Given $G = \langle X \rangle \leq \mathrm{GL}(d, p^e)$, [5] constructs subgroups $H_1, \ldots, H_k$ such that $H_i/S_i \cong T_i$ for some solvable group S_i. Having these H_i at hand, it is possible to construct the permutation group $G/M_3 \leq S_k$, which then can be handled by permutation group methods. Moreover, using the results of [1], [8], the simple groups T_i can be non-constructively recognized.

The Babai–Beals algorithm and its extension by [1], [8] are Monte Carlo, and run in polynomial time in the input length.

Contrary to the geometric approach, [5] does not use the geometry associated with the matrix group action of G. The fact that $G \leq \mathrm{GL}(d, p^e)$ is only used when appealing to a simple consequence of [35], [27]: if T_i is of Lie type in characteristic different from p, then T_i has a permutation representation of degree polynomial in d.

These results can be extended significantly further.

Theorem 3.5 ([34]). *Given $G = \langle X \rangle \leq \mathrm{GL}(d, p^e)$, there is a Las Vegas algorithm that computes the following.*

(i) *The order of G.*

(ii) *A series of subgroups $1 = N_0 \lhd N_1 \lhd \cdots \lhd N_{m-1} \lhd N_m = G$, where N_i/N_{i-1} is a nonabelian simple group or a cyclic group for all i.*

(iii) *A presentation of G.*

(iv) *Given any $g \in \mathrm{GL}(d, p^e)$, the decision whether $g \in G$, and if $g \in G$, then a straight-line program from X, reaching g.*

The algorithm uses an oracle to solve version (DL2) of the discrete logarithm problem in fields of characteristic p and size up to p^{ed}. In the case when all composition factors of Lie type that are in characteristic p are CRDLO, the running time

is polynomial in the input length $|X|d^2e \log p$, plus the time requirement of polynomially many calls to the discrete logarithm oracle. Note that in (ii) the cyclic factors N_i/N_{i-1} may not be simple of prime order because we do not assume that we can factor large integers.

The proof proceeds by continuing the Babai–Beals algorithm when that approach bottoms out. The key idea is that, using the notation introduced in the discussion before Theorem 3.5, for those simple groups T_i that are of Lie type of characteristic p, H_i can be written in an appropriate basis in an upper triangular 3×3 block matrix form such that T_i acts in a quasisimple representation on the block $(2, 2)$. Hence the CRLDO constructive recognition algorithms can be applied. Finally, the subgroup M_1 is handled using a modification of Luks's deterministic algorithm [40] for solvable matrix groups.

The algorithms of [5] and [34] are not practical. However, [5] is a cookie jar of new ideas, which should be used in implementations. Hence, we recently started a project of designing new reduction algorithms for those Aschbacher categories where the current algorithms do not have fast asymptotic running time, combining geometric and black-box methods. This is the second level of unification mentioned in the introduction. Although this project is quite new and there is only one paper [18] in print (about the category of normalizers of extraspecial groups), algorithms for three other categories (imprimitive, tensor product, and tensor induced) are in the offing.

4. A new data structure

In this section we discuss an implementation aspect of computations with permutation and matrix groups. As we have seen in the previous sections, the asymptotically most efficient permutation group algorithms and all existing matrix group algorithms break up the input into manageable pieces. Hence, in order to implement these algorithms, we need a data structure for a recursive scheme which facilitates divide-and-conquer techniques to pass from a group G to a normal subgroup N and the factor group G/N, and then to put together the results of those two smaller computations.

In [42] we describe the design and implementation of such a data structure. This data structure opens up the possibility to handle theoretical algorithms that were considered too complicated for implementation. The homomorphism mechanism is on the black-box level, so permutation groups and matrix groups can be treated in a uniform way. Also, for any group occurring in the recursive scheme, the image of the homomorphism may be either a permutation, matrix, or black-box group, so we can switch between the different types as best suited for the particular application. Once a homomorphism $\varphi\colon H \to K$ is defined for some inner node H of the recursion tree, the computation of $\mathrm{Ker}(\varphi)$ and the combination of the results for $\mathrm{Im}(\varphi)$ and $\mathrm{Ker}(\varphi)$ are done by generic procedures, so in applications we can concentrate on finding suitable homomorphisms φ and the handling of the leaf nodes (which usually means constructive recognition of that node).

The first success story is the implementation of a randomized version of the algorithm of Theorem 2.2. All that was needed was the design and implementation of a randomized speedup for processing groups of Cameron type [36], and an implementation of constructive recognition of alternating and symmetric groups in their natural representation, as described in [48, Section 10.2.4]. The rest of the algorithm of Theorem 2.2 is done automatically by the generic recursive procedure. The final result is the first practical treatment of *all* permutation groups: for small-base inputs, the algorithm reverts to the base-SGS method with minimal overhead (and sometimes it runs faster even on small-base inputs), and on larger inputs there are very substantial savings compared to the straightforward call of the current default base-SGS computation.

The mathematical idea behind the recursion scheme is simple, but there were formidable challenges in the design of the data structure. To pull back the results from N and G/N to G, we need new data types, permutations and matrices with memory, that "remember" how they were obtained from the generators of G by storing a straight-line program (SLP). This results in conflicting requirements. On one hand, these new data types must behave like permutations or matrices, so the *existing and future permutation and matrix group algorithms can be applied to them without rewriting the library of GAP functions* and we can incorporate permutation and matrix group algorithms by other developers who may not need to know of our recursive scheme. On the other hand, the steps of the applied permutation and matrix group algorithms must be recorded in the SLP, although these algorithms are not even aware that this SLP exists. We also need a new type of homomorphism template, flexible enough to accommodate the wide variety of methods that can create factor groups. Exploring an unknown G, we may try a lot of different methods to pass to a factor group; these methods must be prioritized by an automatic method selection mechanism, while at the same time this mechanism must be transparent enough for users to change the order in which applicable methods are called if some extra information is known or suspected about G.

There is a long list of novel tricks incorporated in the new framework: for example, how to balance the recursive tree (that the branches are about the same length, speeding up traversing the tree); how to pass information to the children of a node, so each node can have its own individualized method selection process for a more efficient way to find a homomorphism from this node; and the introduction of an analogue of strong generating sets in the matrix group setting, which enables the writing of shorter straight-line programs to reach group elements. Current work concentrates on adding new homomorphism methods to the recursive scheme for the geometric subgroups in Aschbacher's classification, combining black-box and geometric methods as mentioned at the very end of the previous section, and implementing methods for the end nodes of recursion, to handle almost simple matrix and black-box groups. There will also be methods to handle solvable groups given by power-conjugate presentations, thereby including the third large category of black-box groups in the same scheme. I am quite enthusiastic about this new framework: I think we have found the tool for

the uniform treatment of permutation, matrix, and solvable groups, at the same time unifying the theoretical and practical sides of computations with these groups.

Acknowledgement. I am indebted to Bill Kantor for his very helpful comments.

References

[1] Altseimer, C., Borovik, A. V., Probabilistic recognition of orthogonal and symplectic groups. In *Groups and Computation III* (ed. by W. M. Kantor, Á. Seress), Ohio State Univ. Math. Res. Inst. Publ. 8, Walter de Gruyter, Berlin, New York 2001, 1–20.

[2] Ambrose, S., Neunhöffer, M., Praeger, C. E., Schneider, C., Generalised sifting in black-box groups. *London Math. Soc. J. Comput. Math.* **8** (2005), 217–250.

[3] Aschbacher, M., On the maximal subgroups of the finite classical groups. *Invent. Math.* **76** (1984), 469–514.

[4] Babai, L., Local expansion of vertex-transitive graphs and random generation in finite groups. In *Proc.* 23*rd ACM STOC* 1991, 164–174.

[5] Babai, L., Beals, R., A polynomial-time theory of black box groups. In *Groups St. Andrews* 1997 *in Bath, I* (ed. by C. M. Campbell, E. F. Robertson, N. Ruskuc, G. C. Smith), London Math. Soc. Lecture Note Ser. 260, Cambridge University Press, Cambridge 1999, 30–64.

[6] Babai, L., Cooperman, G., Finkelstein, L., Luks, E. M., Seress, Á., Fast Monte Carlo algorithms for permutation groups. *J. Comput. System Sci.* **50** (1995), 296–308.

[7] Babai, L., Cooperman, G., Finkelstein, L., Seress, Á., Nearly linear time algorithms for permutation groups with a small base. In *Proc. International Symposium on Symbolic and Algebraic Computation* (*ISSAC* '91), ACM Press, New York 1991, 200–209.

[8] Babai, L., Kantor, W. M., Pálfy, P. P., Seress, Á., Black box recognition of finite simple groups of Lie type by statistics of element orders. *J. Group Theory* **5** (2002), 383–401.

[9] Babai, L., Luks, E. M., Seress, Á., Permutation groups in NC. In *Proc. 19th ACM STOC 1987*, 409–420.

[10] Babai, L., Luks, E. M., Seress, Á., Fast management of permutation groups I. *SIAM J. Algorithms* **26** (1997), 1310–1342.

[11] L. Babai, L., Szemerédi, E., On the complexity of matrix group problems I. In *Proc.* 25*th IEEE FOCS* 1984, 229–240.

[12] Beals, R., Babai, L., Las Vegas algorithms for matrix groups. In *Proc.* 34*th IEEE FOCS* 1993, 427–436.

[13] Beals, R., Leedham-Green, C. R., Niemeyer, A. C., Praeger, C. E., Seress, Á., A black-box group algorithm for recognizing finite symmetric and alternating groups, I. *Trans. Amer. Math. Soc.* **355** (2003), 2097–2113.

[14] Bratus, S., Pak, I., Fast constructive recognition of a black-box group isomorphic to S_n or A_n using Goldbach's conjecture. *J. Symbolic Comput.* **29** (2000), 33–57.

[15] Brooksbank, P. A., Fast constructive recognition of black box unitary groups. *London Math. Soc. J. Comput. Math.* **6** (2003), 162–197.

[16] Brooksbank, P. A., Kantor, W. M., On constructive recognition of a black box PSL(d, q). In *Groups and Computation III* (ed. by W. M. Kantor, Á. Seress), Ohio State Univ. Math. Res. Inst. Publ. 8, Walter de Gruyter, Berlin, New York 2001, 95–111.

[17] Brooksbank, P. A., Kantor, W. M., Fast constructive recognition of black box orthogonal groups. Preprint, 2006.

[18] Brooksbank, P. A., Niemeyer, A. C., Seress, Á., A reduction algorithm for matrix groups with an extraspecial normal subgroup. In *Finite Geometries, Groups, and Computation* (ed. by A. Hulpke, R. Liebler, T. Penttila, Á. Seress), Walter de Gruyter, Berlin, New York 2006, 1–16.

[19] Bosma, W., Cannon, J., Playoust, C., The Magma algebra system I: The user language. *J. Symbolic Comput.* **24** (1997), 235–265.

[20] Cameron, P. J., Finite permutation groups and finite simple groups. *Bull. London Math Soc.* **13** (1981), 1–22.

[21] Celler, F., Leedham-Green, C. R., A non-constructive recognition algorithm for the special linear and other classical groups. In *Groups and Computation II* (ed. by L. Finkelstein, W. M. Kantor), Amer. Math. Soc. DIMACS Ser. 28, Amer. Math. Soc., Providence, RI, 1997, 61–67.

[22] Celler, F., Leedham-Green, C. R., Murray, S. H., Niemeyer, A. C., O'Brien, E. A., Generating random elements of a finite group. *Comm. Algebra* **23** (1995), 4931–4948.

[23] Cooperman, G., Finkelstein, L., Combinatorial tools for computational group theory. In *Groups and Computation* (ed. by L. Finkelstein, W. M. Kantor), Amer. Math. Soc. DIMACS Ser. 11, Amer. Math. Soc., Providence, RI, 1993, 53–86.

[24] Cooperman, G., Finkelstein, L., Linton, S., Recognizing $GL_n(2)$ in non-standard representation. In *Groups and Computation II* (ed. by L. Finkelstein, W. M. Kantor), Amer. Math. Soc. DIMACS Series 28, Amer. Math. Soc., Providence, RI, 1997, 85–100.

[25] Conder, M. D. E., Leedham-Green, C. R., Fast recognition of classical groups over large fields. In *Groups and Computation III* (ed. by W. M. Kantor, Á. Seress), Ohio State Univ. Math. Res. Inst. Publ. 8, Walter de Gruyter, Berlin, New York 2001, 113–121.

[26] Conder, M. D. E., Leedham-Green, C. R., O'Brien, E. A., Constructive recognition of PSL(2, q). *Trans. Amer. Math. Soc.* **358** (2006), 1203–1221.

[27] Feit, W., Tits, J., Projective representation of minimum degree of group extensions. *Canad. J. Math.* **30** (1978), 1092–1102.

[28] The GAP Group, GAP – Groups, Algorithms, and Programming, Version 4.4. Aachen–St Andrews 2005; http://www.gap-system.org.

[29] Holt, D., Eick, B., O'Brien, E. A., *Handbook of Computational Group Theory*. Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[30] Kantor, W. M., Simple groups in computational group theory. *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 77–86.

[31] Kantor, W. M., Magaard, K., Black-box exceptional groups of Lie type. In preparation.

[32] Kantor, W. M., Seress, Á., Black box classical groups. *Mem. Amer. Math. Soc.* **149** (2001), Nr. 708.

[33] Kantor, W. M., Seress, Á., Permutation group algorithms via black box recognition algorithms. In *Groups St. Andrews* 1997 *in Bath, II* (ed. by C. M. Campbell, E. F. Robertson, N. Ruskuc, G. C. Smith), London Math. Soc. Lecture Note Ser. 261, Cambridge University Press, Cambridge 1999, 436–446.

[34] Kantor, W. M., Seress, Á., Computing with matrix groups. In *Groups, Combinatorics, and Geometry* (ed. by A. A. Ivanov, M. W. Liebeck, J. Saxl), World Scientific, River Edge, NJ, 2003, 123–137.

[35] Landazuri, V., Seitz, G. M., On the minimal degrees of projective representations of the finite Chevalley groups. *J. Algebra* **32** (1974), 418–443.

[36] Law, M., Niemeyer, A. C., Praeger, C. E., Seress, Á., A reduction algorithm for large-base primitive permutation groups. *London Math. Soc. J. Comput. Math.* **9** (2006), 159–173.

[37] Leedham-Green, C. R., The computational matrix group project. In *Groups and Computation III* (ed. by W. M. Kantor, Á. Seress), Ohio State Univ. Math. Res. Inst. Publ. 8, Walter de Gruyter, Berlin, New York 2001, 229–247.

[38] Liebeck, M. W., O'Brien, E. A., Finding the characteristic of a group of Lie type. Submitted, 2005.

[39] Lovorn, R., Rigorous, subexponential algorithms for discrete logarithms over finite fields. Ph. D. thesis, U. of Georgia, 1992.

[40] Luks, E. M., Computing in solvable matrix groups. In *Proc. 33rd IEEE FOCS*. 1992, 111–120.

[41] Neumann, P. M., Praeger, C. E., A recognition algorithm for special linear groups. *Proc. London Math. Soc.* (3) **65** (1992), 555–603.

[42] Neunhöffer, M., Seress, Á., A data structure for a uniform approach to computations with finite groups. In *Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC'06)*, ACM Press, New York 2006, 254–261.

[43] Niemeyer, A. C., Praeger, C. E., A recognition algorithm for classical groups over finite fields. *Proc. London Math. Soc.* (3) **77** (1998), 117–169.

[44] O'Brien, E. A., Towards effective algorithms for linear groups. In *Finite Geometries, Groups, and Computation* (ed. by A. Hulpke, R. Liebler, T. Penttila, Á. Seress), Walter de Gruyter, Berlin, New York 2006, 163–190.

[45] Ryba, A., Computer construction of split Cartan subalgebras. *J. Algebra*, to appear.

[46] Ryba, A., Identification of matrix generators of a Chevalley group. Submitted, 2005.

[47] Seress, Á., An introduction to Computational Group Theory. *Notices Amer. Math. Soc.* **46** (1997), 671–679.

[48] Seress, Á., *Permutation Group Algorithms*. Cambridge Tracts in Math. 152, Cambridge University Press, Cambridge 2003.

[49] Seress, Á., Large element orders and the characteristic of Lie-type simple groups. Submitted, 2005.

[50] Sims, C. C., Computation with permutation groups. In *Proc. Second Symp. on Symbolic and Algebraic Manipulation*, ACM Press, New York 1971, 23–28.

[51] Sims, C. C., *Computation with Finitely Presented Groups*. Encyclopedia Math. Appl. 48, Cambridge University Press, Cambridge 1994.

Department of Mathematics, The Ohio State University, Columbus, OH 43210, U.S.A.

E-mail: akos@math.ohio-state.edu

Some results in noncommutative ring theory

Agata Smoktunowicz

Abstract. In this paper we survey some results on the structure of noncommutative rings. We focus particularly on nil rings, Jacobson radical rings and rings with finite Gelfand–Kirillov dimension.

Mathematics Subject Classification (2000). 16-02, 16-06, 16N40, 16N20, 16N60, 16D60, 16P90.

Keywords. Nil rings, Jacobson radical, algebraic algebras, prime algebras, growth of algebras, the Gelfand–Kirillov dimension.

1. Introduction

We present here a brief outline of results and examples related mainly to noncommutative nil rings. In this exposition rings are noncommutative and associative. A vector space R is called an algebra (or a K-algebra) if R is equipped with a binary operation

$$* : (R, R) \to R,$$

called multiplication, such that for any $a, b, c \in R$ and for any $\alpha \in K$, we have $(a + b) * c = a * c + b * c$, $a * (b + c) = a * b + a * c$, $(a * b) * c = a * (b * c)$, $\alpha(a * b) = (\alpha a) * b = a * (\alpha b)$.

It is known that simple artinian rings, commutative simple rings and simple right noetherian rings of characteristic zero have unity elements [35]. In this text, rings are usually without 1. In fact nil rings and Jacobson radical rings cannot have unity elements.

2. Nil rings

The most important question in this area is the Köthe Conjecture, first posed in 1930. Köthe conjectured that a ring R with no nonzero nil (two-sided) ideals would also have no nonzero nil one-sided ideals, [24], see also [15] and [27]. This conjecture is still open despite the attention of many noncommutative algebraists. It is a basic question concerning the structure of rings.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The truth of the conjecture has been established for many classes of rings: typically, one proves that for a given class of rings, the sum of all nil one-sided ideals is nil. The most famous examples of such results are the proof of the conjecture in the case of algebras over uncountable fields by Amitsur, and the fact that nil ideals are nilpotent in the class of noetherian rings, proved by Levitski, see [27]. However, as indicated above, Köthe's conjecture is still open in the general case.

An element r in a ring R is said to be *nilpotent* if $r^n = 0$ for some n. A ring R is a *nil* ring if every element of R is nilpotent, and the ring R is *nilpotent* if $R^n = 0$ for some n. A more appropriate definition in the case of infinitely generated rings is the following. A ring R is *locally nilpotent* if every finitely generated subring of R is nilpotent. A thorough understanding of nil and nilpotent rings is important for an attempt to understanding general rings.

In addition, nil rings have some applications in group theory. The following famous theorem was proved in 1964 by Golod and Shafarevich. *For every field F there exists a finitely generated nil F-algebra R which is not nilpotent* ([20]). Recall that a group G is said to be torsion (or periodic) if every $g \in G$ has a finite order. Golod used the group $1 + R$, when F has positive characteristic, to get a counterexample to the General Burnside Problem: *Let G be a finitely generated torsion group. Is G necessarily finite*?

There are many open questions concerning nil rings. As mentioned before, the most important is now known as the *Köthe Conjecture* and was posed by Köthe in 1930: if a ring R has no nonzero nil ideals, does it follow that R has no nonzero nil one-sided ideals? Köthe himself conjectured that the answer would be in the affirmative ([24], [27], [37]).

There are many assertions equivalent to the Köthe Conjecture: For example, the following are equivalent to Köthe's conjecture:

1. The sum of two right nil ideals in any ring is nil.

2. (Krempa [26]) For every nil ring R the ring of 2 by 2 matrices over R is nil.

3. (Fisher, Krempa [18]) For every ring R, R^G is nil implies R is nil (G is the group of automorphisms of R, R^G the set of G-fixed elements).

4. (Ferrero, Puczylowski [17]) Every ring which is a sum of a nilpotent subring and a nil subring must be nil.

5. (Krempa [26]) For every nil ring R the polynomial ring $R[x]$ in one indeterminate over R is Jacobson radical.

6. (Smoktunowicz [44]) For every nil ring R the polynomial ring $R[x]$ in one indeterminate over R is not left primitive.

7. (Xu [49]) The left annihilators of a single element in every complement of a nil radical in a maximal left nil ideal satisfy a.c.c.

Recall that a ring R is Jacobson radical if for every $r \in R$ there is $r' \in R$ such that $r + r' + rr' = 0$. Every nil ring is Jacobson radical. The largest ideal in a ring R, which is Jacobson radical is called the *Jacobson radical* of R. The Jacobson radical of a ring R equals the intersection of all (right) primitive ideals of R (I is a primitive ideal in R if I/R is primitive). Recall that a ring R is (right) primitive if there is a maximal right ideal Q such that $Q + I = R$ for every nonzero ideal I in R and there is $b \in R$ such that $br - r \in Q$ for every $r \in R$ ([13]).

The Köthe Conjecture is said to hold for a ring R if the ideal generated by the nil left ideals of R is nil. Köthe's conjecture holds for the class of Noetherian rings (Levitzki, [27], [32]), Goldie rings (Levitzki, [32]), rings with right Krull dimension (Lenagan [29], [15]), monomial algebras (Beidar, Fong [6]), PI rings (Rasmyslow–Kemer–Braun [14], [34], [22], [12]), algebras over uncountable fields (Amitsur [27], [36]).

There are many related results, some are indicated in the following.

Theorem 2.1 (Levitzki; [32]). *Let R be a right Noetherian ring. Then every nil one-sided ideal of R is nilpotent.*

Theorem 2.2 (Lenagan [29]). *If R has right Krull dimension, then nil subrings of R are nilpotent.*

Theorem 2.3 (Gordon, Lenagan and Robson, Gordon and Robson; [15]). *If R has right Krull dimension, then the prime radical of R is nilpotent.*

The prime radical of R is a nil ideal and is equal to the intersection of all prime ideals in R.

Theorem 2.4 (Beidar, Fong [6]). *Let X be a nonempty set, $Z = \langle X \rangle$ the free monoid on X, Y an ideal of the monoid Z, and F a field. Then the Jacobson radical of the monomial algebra $F[Z/Y]$ is locally nilpotent.*

In the case of characteristic zero the result is due to Jaspers and Puczylowski, [21]. Earlier, Belov and Gateva-Ivanova [10] showed that the Jacobson radical of a finitely generated monomial algebra over a field is nil. However, it is not true that the Jacobson radical of a finitely generated monomial algebra is nilpotent, since it was shown by Zelmanov [50] that there is a finitely generated prime monomial algebra with a nonzero locally nilpotent ideal.

Theorem 2.5 (Razmyslov–Kemer–Braun [34], [22], [12]; [14]). *If R is a finitely generated PI-algebra over a field then the Jacobson radical of R is nilpotent.*

Razmyslov [34] proved this for rings satisfying all identities of matrices, Kemer [22] for algebras over fields of characteristic zero. Later Braun [12] proved the nilpotency of the radical in any finitely generated PI algebra over a commutative noetherian ring. Amitsur has previously shown that the Jacobson radical of a finitely generated PI algebra over a field is nil.

Another famous result is the Nagata–Higman Theorem:

Theorem 2.6 (Nagata–Higman; [19]). *If A is an associative algebra of characteristic p such that $a^n = 0$ for all $a \in A$ and $p > n$ or $p = 0$ then A is nilpotent.*

For interesting results related to Nagata–Higman's theorem see [19].

A theorem of Klein [23] asserts that if R is a nil ring of bounded index then $R[x]$ is a nil ring of bounded index.

In 1956 Amitsur [27] showed that if R is a nil algebra over an uncountable field, then the polynomial ring $R[x]$ in one indeterminate over R is also nil. The situation is completely different for countable fields, as was shown by the author in 2000.

Theorem 2.7 (Smoktunowicz [43]). *For every countable field K there is a nil K-algebra N such that the polynomial ring in one indeterminate over N is not nil.*

This answers a question of Amitsur. Another important theorem by Amitsur is the following.

Theorem 2.8 (Amitsur; [27]). *Let R be a ring. Then the Jacobson radical of the polynomial ring $R[x]$ is equal to $N[x]$ for some nil ideal N of R.*

In 1956 Amitsur conjectured that if R is a ring, and $R[x]$ has no nil ideals then it is semiprimitive (i.e. the Jacobson radical of $R[x]$ is zero). This assertion is true for many important classes of rings, as mentioned above. However, the following theorem shows that this conjecture does not hold in general: *There is a nil ring N such that the polynomial ring in one indeterminate over N is Jacobson radical but not nil* ([41]). For some generalizations of this theorem see [45]. This theorem is true in a more general setting: *For every natural number n, there is a nil ring N such that the polynomial ring in n commuting indeterminates over N is Jacobson radical but not nil*.

Recall that, as shown by Krempa in [26], Köthe's conjecture is equivalent to the assertion that polynomial rings over nil rings are Jacobson radical. However, homomorphic images of polynomial rings over nil rings with nonzero kernels are often Jacobson radical, as is shown by the next result.

Theorem 2.9 (Smoktunowicz [44]). *Let R be a nil ring and $R[x]$ the polynomial ring in one indeterminate over R. Let I be an ideal in $R[x]$ and M the ideal of R generated by coefficients of polynomials from I. Then $R[x]/I$ is Jacobson radical if and only if $R[x]/M[x]$ is Jacobson radical.*

The following are interesting open questions on nil rings.

Question 1 (Latyshev, [16], pp. 12). Let A be an associative algebra with a finite number of generators and relations. If A is a nil algebra must it be nilpotent?

Question 2 (Amitsur; [33]). Let A be an associative algebra with a finite number of generators and relations. Does it follow that the Jacobson radical of A nil?

3. Algebraic algebras

The most well-known question in this area is the Kurosh Problem ([15], [36]). *Let R be a finitely generated algebra over a field F such that R is algebraic over F. Is R finite dimensional over F?*

This problem has a negative solution in general. The famous construction of Golod and Shafarevich in the 1960s produced a finitely generated nil algebra which is not nilpotent ([20]). This was then used to construct a counterexample to the Burnside Conjecture, one of the biggest outstanding problems in group theory at that time. Zelmanov was later awarded the Fields Medal for his solution of the Restricted Burnside Problem [27].

However, the Kurosh Problem is still open for the key special case of a division ring:

Question 3 (Kurosh's problem for division rings [16], [36]). Let R be a finitely generated algebra over a field F such that R is algebraic over F and R is a division ring. Does it follow that R a finite dimensional vector space over its center?

Again, as with the nil ring problems, there are many partial results. The Kurosh Problem for division rings is still open in general, but it is answered affirmatively for F finite and for F having only finite algebraic field extensions, in particular, for F algebraically closed ([36]). By Levitzki's and Kaplanski's theorem, Kurosh's conjecture is also true if there is a bound on the degree of elements in R ([15]). It is unknown whether Kurosh's problem for division rings has a positive answer in the case of algebras over uncountable fields. Also the following question is still open: Is Kurosh's conjecture true for division rings with finite Gelfand–Kirillov dimension, and in particular for division rings with quadratic growth? There are obvious connections with problems in nil rings. A nil element is obviously algebraic, and, in the converse direction, it is possible to construct an associated graded algebra connected with an algebraic algebra in such a way that the positive part is graded nil, i.e., all homogeneous elements are nil. On the other hand, the Kurosh Problem has a negative solution for rings with finite Gelfand–Kirillov dimension ([30]), for simple rings ([42]), for primitive rings ([2]), for finitely generated primitive rings ([8]), and for finitely generated algebraic primitive rings ([9]). However, a natural question arising from the general Kurosh Problem remains open:

Question 4 (Small's question). Let R be a finitely generated simple algebra with 1 over a field F such that R is algebraic over F. Is R a finite dimensional vector space over its center?

Another open question on division rings, which has been around for years, is the following:

Question 5. Let K be a field and let R be a finitely generated algebra which is a division ring. Does it follow that R is a finitely generated vector space over K?

As far as I know this question is very much open even with various conditions, like e.g. Gelfand–Kirillov dimension 2. It has been shown by Small ([38]) that a division ring which is a homomorphic image of a graded noetherian ring (of course, by a non graded ideal) must be finite dimensional. There is a similar open question concerning rings:

Question 6 ([16], p. 20). Does there exist an infinite associative division ring which is finitely generated as a ring?

4. Algebras with finite Gelfand–Kirillov dimension

The Gelfand–Kirillov dimension measures the rate at which an algebra is generated by a generating set. The GK dimension is zero for algebras which are finite dimensional and an elementary counting argument shows that the next possible dimension is one. However, Borho and Kraft showed that any real number value greater than or equal to two is possible ([25]). Bergman's famous Gap Theorem establishes that there is no algebra with GK dimension strictly between one and two ([11], see also [25]). A theorem of Small and Warfield asserts that an affine prime algebra R over a field F of GK dimension 1 is a finite module over its center, and that its center is a finitely generated F-algebra of GK dimension 1 ([40], [25]). In the special case when R is a finitely generated domain over an algebraically closed field with GK dimension 1, it follows by Small–Warfield's and Tsen's theorem (see [15]) that R is in fact commutative ([47]). A theorem of Small, Stafford and Warfield shows that a finitely generated algebra with GK dimension 1 is close to being commutative in that it must satisfy a polynomial identity ([39], [25]).

The graded case has attracted interest in the last decade or so with the development of noncommutative algebraic geometry. Here progress is being made by studying algebras with restricted conditions, including conditions on the growth of the algebras. Low GK dimension examples are obviously of interest. Since the theory is developing by analogy with the classical projective case, one typically deals with graded algebras. Thus dimensions should be increased by one compared to the ungraded case. The first interesting case is to study graded domains of GK dimension two; that is, noncommutative projective curves. This was done in a famous paper in the *Inventiones Mathematicae* by Artin and Stafford about 10 years ago. In fact Artin and Stafford described in [3] the structure of finitely graded domains in terms of algebras related to automorphisms of elliptic curves. They were able to tell when such algebras are noetherian, primitive, PI, etc. In this paper they formulated the analogue of the Bergman Gap Theorem: there should be no graded (by natural numbers) domain with GK dimension strictly between two and three, and they were able to exclude the open interval $(2, 11/5)$. The author has recently established in [46] the truth of the full conjecture.

There are several connecting threads between the three areas mentioned above. As stated earlier, nil elements are algebraic, and graded nil algebras can be constructed from algebraic algebras as associated graded rings. The Golod–Shafarevich construction yields a nil but *not* nilpotent algebra which has exponential growth and so certainly infinite GK dimension.

In recent work with Lenagan, the author has constructed an example of a finitely generated nil but not nilpotent algebra that has finite GK dimension (≤ 20). The precise growth condition dividing nilpotent and nil but not nilpotent is tantalizing. Certainly, nil algebras with GK dimension 1 are easily seen to be nilpotent. It may be that the dividing line is of quadratic growth.

In this area the following question remains open and may be considered to be a test question for new methods. Is there a finitely generated nil algebra with quadratic growth which is not nilpotent? An F-algebra R has quadratic growth if there is a constant c and a generating subspace V of R such that $\dim_F(V+V^2+\cdots+V_n) < cn^2$ for all $n > 0$. In particular GKdim $R \leq 2$.

In connection with this problem, a recent result of Bartholdi is pertinent. In 2004 Bartholdi proved the following result.

Theorem 4.1 (Bartholdi [4]). *Let K be an algebraic field extension of F_2. Then there exist a finitely generated graded K-algebra R such that all homogeneous elements of R are nil, but the algebra has a transcendental invertible element. In particular, R is graded nil but not nil. This algebra R has also a subalgebra isomorphic to the ring of 2×2 matrices over R.*

In more detail, Bartholdi showed that an affine 'recurrent transitive' algebra (without unit) constructed from Grigorchuk's group of intermediate growth is of quadratic growth. Moreover, assuming that the base field is an algebraic extension of F_2, the algebra is Jacobson radical and not nil. This algebra R was earlier studied by Ana Christina Vieira in [48], who showed that R is prime and for every non-zero two sided ideal I of R, R/I is finite-dimensional.

Another way to construct examples of finitely generated algebras was introduced by Markov and later extended by Beidar ([5]), Bell and Small ([7], [8], [32], [36]). The effect of Markov's result is to allow constructions first in infinitely generated algebras, thus simplifying the problem, and then, by using Markov's method, to bring the construction into a finitely generated algebra.

Theorem 4.2 (Markov [31]). *Let K be a field, and let R be a prime, countably generated K-algebra. Then there exists a prime K-algebra A generated by two elements x, y such that R is isomorphic to a right ideal of A, namely to xR.*

Recall that $T \subseteq R$ is a corner of an algebra R if T is a subalgebra of R and $TRT \subseteq T$. Markov's theorem was extended by Small who showed (around 1982, unpublished) that if K is a field and T is a prime, countably generated K algebra then there exists a finitely generated, prime K-algebra A such that T is a corner of A.

It is possible to apply this result in many situations. For example, in [8], Bell and Small applied the result to show that there is a finitely generated algebraic primitive algebra which is infinitely dimensional over its center. In 2003 Bell proved the following extension of Small's theorem. *Let K be a field, and let T be a prime, countably generated K-algebra of Gelfand–Kirillov dimension $\alpha < \infty$. Then there exists a finitely generated, prime K-algebra A of Gelfand–Kirillov dimension $\alpha + 2$ such that T is a corner of A* (see [7]).

Bell's theorem above is related to another question of Small: if R is a noetherian affine algebra with quadratic growth, does it follow that R is either primitive or PI? This is true in the graded case, as was shown by Artin and Stafford in 2000. According to Small, it is also true if every non-zero prime ideal in R is maximal.

An application by Bell of his theorem is the following example which is a counterexample to another question of Small. There is a prime, affine algebra with Gelfand–Kirillov dimension 2 which is not PI and not primitive. This algebra has a nonzero Jacobson radical. The following result of Lanski, Resco and Small assures that usually an affinization of a primitive ring is still primitive:

Theorem 4.3 (Lanski, Resco, Small [28]). *Let R be a prime ring. Then the following is true:*

1. *Let V be a right ideal of R. Then R is a primitive ring exactly when $V/(V \cap l(V))$ is a primitive ring, where $l(V) = \{r \in R : rV = 0\}$.*
2. *If R contains an idempotent e, then R is a primitive ring if and only if eRe is a primitive ring.*

5. Simple rings

A ring R (possibly without 1) is called simple if $R^2 \neq 0$ and R has no proper two-sided ideals. Levitzki, Jacobson, Kaplansky and others asked if there is a simple nil ring. An example of a simple ring which is a Jacobson radical ring (that is, $R = J(R)$ where $J(R)$ denotes the Jacobson radical of R) was found by Sasiada in 1961, see e.g. [15]; however, this ring is not nil. Note that the polynomial ring in one indeterminate over Sasiada's ring is left and right primitive ([44]). By Nakayama's lemma a simple Jacobson radical ring cannot be finitely generated. Since every nil ring is Jacobson radical, a simple nil ring also cannot be finitely generated. A few years ago examples of simple nil rings were constructed by the author ([15]).

Theorem 5.1 (Smoktunowicz [42]). *For every countable field K there is a simple nil algebra over K.*

Notice that all rings in that paper were graded by integers. The following natural question remains open.

Question 7. Is there a simple noncommutative nil algebra over an uncountable field?

Acknowledgements. The author is very grateful to Tom Lenagan and Lance Small for many useful suggestions, and to Tom Lenagan for his collaboration in writing Section 4.

References

[1] Amitsur, S. A., A generalization of Hilbert's Nullstellensatz. *Proc. Amer. Math. Soc.* **8** (1957), 649–656.

[2] Amitsur, S. A., unpublished.

[3] Artin, M., Stafford, J. T., Noncommutative graded domains with quadratic growth. *Inventiones Math*. **122** (1995), 231–276.

[4] Bartholdi, L., Branch Rings, thinned rings, tree enveloping rings. *Israel J. Math.*, to appear.

[5] Beidar, K. I., Radicals of finitely generated algebras. *Uspiekhi Mat. Nauk.* **222** (1981), 203–204.

[6] Beidar, K. I., Fong, Y., On radicals of monomial algebras. *Comm. Algebra* **26** (1998), 3913–3919.

[7] Bell, J., Examples in finite Gelfand-Kirillov dimension. *J. Algebra* **263** (2003), 159–175.

[8] Bell, J., Small, L. W., A question of Kaplansky. *J. Algebra* **258** (2002), 386–388.

[9] Bell, J., Lenagan, T., Small, L., Smoktunowicz, A., unpublished.

[10] Belov, A., Gateva-Ivanova, T., Radicals of monomial algebras. In *First International Tainan–Moscow Algebra Workshop* (Tainan 1994), De Gruyter, Berlin 1996, 159–169.

[11] Bergman, G. M., *A note of growth functions of algebras and semigroups*. Mimeographed notes, University of California, Berkeley 1978.

[12] Braun, A., The nilpotency of the radical in a finitely generated PI ring. *J. Algebra* **89** (1984), 375–396.

[13] Divinski, N. J., *Rings and radicals*. Mathematical Expositions No.14, University of Toronto Press, Toronto, Ont., 1965.

[14] Drenski, V., *Polynomial identity rings*. Advanced Courses in Mathematics, Barcelona. Birkhäuser, Basel 2004.

[15] Faith, C., *Rings and Things and a Fine Array of Twentieth Century Associative Algebra*. Math. Surveys Monogr. 65, Amer. Math. Soc., Providence, RI, 1999; 2nd ed., 2004.

[16] Filippov, V. T., Kharchenko, V. K., Shestakov, I., P. (eds.), *The Dniester Notebook: Unsolved Problems in Theory of Rings and Modules*. Fourth edition, Mathematics Institute, Russian Academy of Sciences, Siberian Branch, Novosibirsk 1993.

[17] Ferrero, M., Puczylowski, E. R., On rings which are sums of two subrings. *Arch. Math.* (*Basel*) **53** (1989), 4–10.

[18] Fisher, J. W., Krempa, J., "R^G is nil implies R is nil" is equivalent to the "Koethe conjecture". *Houston J. Math.* **9** (1983), 177–180.

[19] Formanek, E., The Nagata-Higman Theorem. *Acta Appl. Math.* **21** (1990), 185–192.

[20] Golod, E. S., Shafarevich, I. R., On the class field tower. *Izv. Akad. Nauk. SSSR Mat. Ser.* **28** (1964), 261–272.

[21] Jespers, E., Puczylowski, E. R., The Jacobson radical and Brown-McCoy radical of rings graded by free groups. *Comm. Algebra* **19** (1991), 551–558.

[22] Kemer, A. R., Capelli identities and nilpotency of the radical of finitely generated PI algebra. *Dokl. Akad. Nauk SSSR* **255** (1980), 739–797.

[23] Klein, A. A., Rings with bounded index of nilpotence. *Contemp. Math.* **13** (1982), 151–154.

[24] Köthe, G., Die Struktur der Ringe, deren Restklassenring nach dem Radikal vollständing reduzibel ist. *Math. Z.* **32** (1930), 161–186.

[25] Krause, G., Lenagan, T. H., *Growth of Algebras and Gelfand-Kirillov Dimension*. Revised edition, Graduate Studies in Mathematics 22, Amer. Math. Soc., Providence, RI, 2000.

[26] Krempa, J., Logical connections between some open problems concerning nil rings. *Fund. Math.* **76** (1972), 121–130.

[27] Lam, T. Y., *A first course in Noncommutative rings*. Second edition, Graduate Texts in Mathematics 131, Springer-Verlag, New York 2001

[28] Lanski, C., Resco, R., Small, L., On the primitivity of prime rings. *J. Algebra* **59** (1979), 395–398.

[29] Lenagan, T. H., The nil radical of a ring with Krull dimension. *Bull. London Math. Soc.* **5** (1973), 307–311.

[30] Lenagan, T. H., Smoktunowicz, A., An infinite dimensional affine nil algebra with finite Gelfand-Kirillov dimension. Submitted.

[31] Markov, V. T., Some examples of finitely generated algebras. *Uspekhi Mat. Nauk* **221** (1981), 185–186.

[32] McConnel, J. C., and Robson, J. C., *Noncommutative Noetherian Rings*. Wiley Interscience, Chichester 1987.

[33] Puczylowski, E. R., Some results and questions on nil rings. In *15th School of Algebra* (Portuguese) (Canela, 1998), *Mat. Contemp.* **16** (1999), 265–280.

[34] Razmyslov, Ju. P., The Jacobson radical in PI algebras. *Algebra i Logika* **13** (1974), 337–360.

[35] Robson, J. C., Do simple rings have unity elements? *J. Algebra* **7** (1967), 140–143.

[36] Rowen, L. H., *Ring Theory*. Vol. I, II, Pure and Applied Mathematics 127, 128, Academic Press, Inc., Boston, MA, 1988.

[37] Rowen, L. H., Koethe's conjecture. In *Ring Theory* 1989 (Ramat Gan and Jerusalem, 1988/1989), Israel Math. Conf. Proc. 1, Weizmann Science Press of Israel, Jerusalem 1989, 193–202.

[38] Small, L. W., private communication.

[39] Small, L. W., Stafford, J. T., Warfield, R. B., Jr, Affine algebras of Gelfand-Kirillov dimension one are PI. *Math. Proc. Cambridge Philos. Soc.* **97** (1984), 407–414.

[40] Small, L. W., Warfield, R. B., Jr, Prime affine algebras of Gelfand-Kirillov dimension one. *J. Algebra* **91** (1984) 384–389.

[41] Smoktunowicz, A., Puczyłowski, E. R., A polynomial ring that is Jacobson radical but not nil. *Israel J. Math*. **124** (2001), 317–325.

[42] Smoktunowicz, A., A simple nil ring exists. *Comm. Algebra* **30** (2002), 27–59.

[43] Smoktunowicz, A., Polynomial rings over nil rings need not be nil. *J. Algebra* **233** (2000), 427–436.

[44] Smoktunowicz, A., On primitive ideals in polynomial rings over nil rings. *Algebr. Represent. Theory* **8** (2005), 69–73.

[45] Smoktunowicz, A., Amitsur's conjecture on polynomial rings in n-commuting indeterminates. *Math. Proc. Roy. Irish Acad.* **102** (2002), 205–213.

[46] Smoktunowicz, A., There are no graded domains with GK dimension strictly between 2 and 3. *Invent. Math*, to appear.

[47] Stafford, J. T., Van den Bergh, M., Noncommutative curves and noncommutative surfaces, *Bull. Amer. Math. Soc.* **38** (2001) 171–216.

[48] Vieira, A. C., Modular algebras of Burnside p-groups. In *16th School of Algebra, Part II* (Portuguese) (Brasília, 2000), *Mat. Contemp.* **21** (2001), 287–304.

[49] Xu, Y. H., On the Koethe problem and the nilpotent problem. *Sci. Sinica Ser. A* **26** (1983), 901–908.

[50] Zelmanov, E. I., An example of a finitely generated prime ring. *Sibirsk. Mat. Zh.* **20** (1979), 303–304.

[51] Zelmanov, E. I., Solution of the restricted Burnside problem for groups of odd exponent. *Izv. Akad. Nauk SSSR Ser. Mat.* **54** (1990) 42–59.

[52] Zelmanov, E. I., Solution of the restricted Burnside problem for 2-groups. *Mat. Sb.* **182** (1991), 568–592.

School of Mathematics, University of Edinburgh, JCMB, King's Buildings,
EH9 3JZ Edinburgh, Scotland, U.K.
and
Institute of Mathematics of the Polish Academy of Sciences, ul. Sniadeckich 8,
P.O. Box 137, 00-950 Warsaw, Poland

E-mail: A.Smoktunowicz@ed.ac.uk A.Smoktunowicz@impan.gov.pl

Higher composition laws and applications

Manjul Bhargava*

Abstract. In 1801 Gauss laid down a remarkable law of composition on integral binary quadratic forms. This discovery, known as *Gauss composition*, not only had a profound influence on elementary number theory but also laid the foundations for ideal theory and modern algebraic number theory. Even today, Gauss composition remains one of the best ways of understanding ideal class groups of quadratic fields.

The question arises as to whether there might exist similar laws of composition on other spaces of forms that could shed light on the structure of other algebraic number rings and fields. In this article we present several such higher analogues of Gauss composition, and we describe how each of these composition laws can be interpreted in terms of ideal classes in appropriate rings of algebraic integers. We also discuss several applications of these composition laws, including the resolution of a critical case of the Cohen–Lenstra–Martinet heuristics, and a solution of the long-standing problem of counting the number of quartic and quintic fields of bounded discriminant. In addition, we describe the mysterious relationship between these various composition laws and the exceptional Lie groups. Finally, we discuss prospects for future work and conclude with several open questions.

Mathematics Subject Classification (2000). Primary 11R29; Secondary 11R45.

Keywords. Gauss composition, classical invariant theory, density theorems.

1. Introduction

Gauss published his seminal treatise *Disquisitiones Arithmeticae* in 1801. One of the primary subjects of this work was the (integral) *binary quadratic form*, i.e., any expression $f(x, y) = ax^2 + bxy + cy^2$ where $a, b, c \in \mathbb{Z}$.[1] The group $\mathrm{SL}_2(\mathbb{Z})$ acts naturally on the space of binary quadratic forms by linear substitution of variable: if $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, then one defines

$$(\gamma \cdot f)(x, y) = f((x, y)\gamma).$$

Gauss studied this action of $\mathrm{SL}_2(\mathbb{Z})$ on binary quadratic forms f in terms of the *discriminant* $\mathrm{Disc}(f) = b^2 - 4ac$, as it is easily seen that this discriminant remains

*The author was partially supported by a Packard Fellowship. I am extremely grateful to Andrew Wiles and Peter Sarnak for their encouragement and to Jonathan Hanke, Wei Ho, and Melanie Wood for numerous helpful comments.

[1] Gauss actually considered only the forms where b is even; however, from the modern point of view it is more natural to assume a, b, c are arbitrary integers.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$. In fact, one can show that *any* polynomial $P(a, b, c)$ invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$ on the space of binary quadratic forms $ax^2 + bxy + cy^2$ must be a polynomial in the discriminant $b^2 - 4ac$ (see e.g. [28]).

It follows that the binary quadratic forms of any fixed discriminant D also naturally break up into orbits under the action of $\mathrm{SL}_2(\mathbb{Z})$. We say a quadratic form $ax^2 + bxy + cy^2$ is *primitive* if a, b, c are relatively prime. Then $\mathrm{SL}_2(\mathbb{Z})$ evidently preserves primitivity, so that the primitive forms of a given discriminant also break up into $\mathrm{SL}_2(\mathbb{Z})$-orbits. Gauss's remarkable discovery regarding these primitive $\mathrm{SL}_2(\mathbb{Z})$-orbits was the following:

Theorem 1.1 (Gauss). *Let $D \equiv 0$ or 1 modulo 4. Then the set of $\mathrm{SL}_2(\mathbb{Z})$-orbits of primitive binary quadratic forms having discriminant D naturally possesses the structure of a finite abelian group.*

What is particularly remarkable about this theorem is that Gauss proved this result before the notion of group formally existed! Theorem 1.1 is quite a deep fact, and has a number of beautiful interpretations. Classically, the theorem generalizes the identity of Brahmagupta [12]:

$$(x_1^2 + Dy_1^2)(x_2^2 + Dy_2^2) = x_3^2 + Dy_3^2,$$

where $x_3 = x_1x_2 + Dy_1y_2$ and $y_3 = x_1y_2 - y_1x_2$. Gauss's theorem describes all identities of the form

$$(a_1x_1^2 + b_1x_1y_1 + c_1y_1^2)(a_2x_2^2 + b_2x_2y_2 + c_2y_2^2) = (a_3x_3^2 + b_3x_3y_3 + c_3y_3^2) \quad (1)$$

where x_3 and y_3 are bilinear functions of (x_1, y_1) and (x_2, y_2) with integer coefficients. Because of the bilinearity condition on (x_3, y_3), the existence of an identity of the type (1) depends only on the $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes of the three forms. Remarkably, the ensemble of all such identities turns the set of $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes of primitive quadratic forms of discriminant D into a group for any eligible value of D. This is precisely the group described by Gauss in Theorem 1.1, showing in particular that the theorem is compatible with the multiplicative structure of the values taken by the forms.

In modern language, the group described in Theorem 1.1 is simply the narrow class group of the unique quadratic ring $S(D)$ of discriminant D (see Section 2.1). This connection with ideal class groups was in fact one of the original motivations for Dedekind to introduce "ideal numbers", or what are now called ideals. Thus Theorem 1.1 really lies at the foundations of modern algebraic number theory. Moreover, Gauss composition still remains one of the best methods for understanding narrow class groups of quadratic fields, and it is certainly still the best way of computing with them.

Of course, Gauss's composition law is related in this way only to field extensions of $\mathbb{Q}$ of degree two, and it would be desirable to have similar ways to understand cubic, quartic, and higher degree fields. The question thus arises: do there exist analogous

composition laws on other spaces of forms, which could be used to shed light on the structure of higher degree fields?

2. The parametrization of algebraic structures

2.1. Gauss composition and rings of rank 2. An alternate way of viewing Gauss composition is as a parametrization result. To describe this, we need some simple definitions. First, define a *ring of rank n* to be any commutative ring with identity whose underlying additive group is isomorphic to $\mathbb{Z}^n$. For example, an order in a number field of degree n is a ring of rank n. Rings of rank 2, 3, 4, 5, and 6 are called *quadratic*, *cubic*, *quartic*, *quintic*, and *sextic* rings respectively. In general, a ring $\mathcal{R}$ of rank n is said to be an *order* in a $\mathbb{Q}$-algebra K if $\mathcal{R} \otimes \mathbb{Q} = K$.

Given a ring $\mathcal{R}$ of rank n, there are two simple functions $\mathcal{R} \to \mathbb{Z}$ called the *trace* and the *norm*, denoted by Tr and N respectively. Given $\alpha \in \mathcal{R}$, we define $\mathrm{Tr}(\alpha)$ (resp. $\mathrm{N}(\alpha)$) as the trace (resp. determinant) of the linear map $\mathcal{R} \xrightarrow{\times\alpha} \mathcal{R}$ given by multiplication by α. The function $x, y \mapsto \mathrm{Tr}(xy)$ defines an inner product on $\mathcal{R}$. If $\langle \alpha_0, \ldots, \alpha_{n-1} \rangle$ is a $\mathbb{Z}$-basis of $\mathcal{R}$, then the *discriminant* $\mathrm{Disc}(\mathcal{R})$ is defined to be the determinant $\mathrm{Det}(\mathrm{Tr}(\alpha_i \alpha_j)_{0 \le i,j \le n-1})$. In basis-free terms, the discriminant of $\mathcal{R}$ is the co-volume of the lattice $\mathcal{R}$ with respect to this inner product, and forms the most important invariant of a ring of rank n. It turns out that the discriminant is always an integer congruent to 0 or 1 (mod 4).

It is easy to describe what all quadratic rings are in terms of the discriminant. Namely, for every integer D congruent to 0 or 1 modulo 4, there is a unique quadratic ring $S(D)$ having discriminant D (up to isomorphism), given by

$$S(D) = \begin{cases} \mathbb{Z}[x]/(x^2) & \text{if } D = 0, \\ \mathbb{Z} \cdot (1,1) + \sqrt{D} \cdot (\mathbb{Z} \oplus \mathbb{Z}) & \text{if } D \ge 1 \text{ is a square}, \\ \mathbb{Z}[(D + \sqrt{D})/2] & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, if we denote by $\mathbb{D}$ the set of elements of $\mathbb{Z}$ that are congruent to 0 or 1 (mod 4), we may say that isomorphism classes of quadratic rings are parametrized by $\mathbb{D}$. The case $D = 0$ is called the *degenerate* case.

Gauss composition concerns the parametrization of narrow (or oriented) ideal classes in oriented quadratic rings. An *oriented* quadratic ring is a quadratic ring in which one of the two choices for a square root $\sqrt{D}$ of D has been distinguished, where D denotes the discriminant of the ring.[2] An *oriented* ideal of a nondegenerate quadratic ring S is a pair (I, ε), where I is any ideal of S having rank 2 over $\mathbb{Z}$ and $\varepsilon = \pm 1$ gives the *orientation* of I. Multiplication of oriented ideals is defined

[2]The advantage of this point of view is that any two oriented quadratic rings of the same discriminant are then canonically isomorphic; to construct this isomorphism, one simply sends the distinguished $\sqrt{D}$ in one ring to that in the other. Note that a choice of $\sqrt{D}$ amounts to a choice of generator of $\wedge^2 S$, namely $1 \wedge \left(\frac{D+\sqrt{D}}{2}\right)$ – hence the name *oriented* quadratic ring.

componentwise. Similarly, for an element $\kappa \in K = S \otimes \mathbb{Q}$, the product $\kappa \cdot (I, \varepsilon)$ is defined to be the ideal $(\kappa I, \mathrm{sgn}(N(\kappa))\varepsilon)$. Two oriented ideals (I_1, ε_1) and (I_2, ε_2) are said to be in the same *class* if $(I_1, \varepsilon_1) = \kappa \cdot (I_2, \varepsilon_2)$ for some invertible $\kappa \in K$. In practice, we will denote an oriented ideal (I, ε) simply by I, with the orientation $\varepsilon = \varepsilon(I)$ on I being understood.

In this language, Gauss composition states:

Theorem 2.1. *There is a canonical bijection between the set of* $\mathrm{SL}_2(\mathbb{Z})$*-equivalence classes of nondegenerate binary quadratic forms and the set of isomorphism classes of pairs* (S, I)*, where* S *is a nondegenerate oriented quadratic ring and* I *is an oriented ideal class of* S*.*

The map from oriented ideal classes to binary quadratic forms is easily described. Given an oriented ideal $I \subset S$, let $\langle \alpha_1, \alpha_2 \rangle$ be a correctly oriented basis of I, i.e., a basis such that the determinant of the change-of-basis matrix from $\langle 1, \sqrt{D} \rangle$ to $\langle \alpha_1, \alpha_2 \rangle$ has the same sign as $\varepsilon(I)$;[3] this determinant is called the *norm* of I and is denoted $\mathrm{N}(I)$. To the oriented ideal I, one then associates the binary quadratic form

$$Q(x, y) = \frac{\mathrm{N}(\alpha_1 x + \alpha_2 y)}{\mathrm{N}(I)}. \tag{3}$$

One readily verifies that $Q(x, y)$ is an integral binary quadratic form and that it is well-defined up to the action of $\mathrm{SL}_2(\mathbb{Z})$. What is remarkable about Theorem 2.1 is not just that every oriented ideal class of a quadratic ring yields an integral binary quadratic form, but that every integral binary quadratic form arises in this way! Another remarkable aspect of the correspondence (3) of Theorem 2.1 is that it is *discriminant-preserving*: under the bijection, the discriminant of a binary quadratic form is equal to the discriminant of the corresponding quadratic ring. That is, oriented ideal classes in $S(D)$ correspond to $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes of binary quadratic forms having discriminant D.

An oriented ideal I of the oriented quadratic ring $S(D)$ is said to be *invertible* if there exists a (fractional) oriented ideal I' such that the product $I I'$ is $(S(D), +1)$. It is known that the set of invertible oriented ideals modulo multiplication by scalars forms a finite abelian group $\mathrm{Cl}^+(S(D))$, called the *oriented* (or *narrow*) *class group*.[4] One checks that invertible oriented ideals correspond to primitive forms via (3). Gauss's group structure on classes of primitive forms of discriminant D arises from the fact that the invertible oriented ideal classes of a quadratic ring $S(D)$ form a group under multiplication.

In the statement of the theorem, we have used the word "nondegenerate" to mean "nonzero discriminant". Theorem 2.1 could also be extended to zero discriminant, although this would require a rather more involved notion of "oriented ideal class", so in what follows we always restrict ourselves to the nondegenerate case.

[3] Evidently, for any basis $\langle \alpha_1, \alpha_2 \rangle$ of I, either $\langle \alpha_1, \alpha_2 \rangle$ or $\langle \alpha_2, \alpha_1 \rangle$ will be correctly oriented.

[4] The usual *class group* is a quotient of the oriented class group, and may be obtained by "forgetting" all orientations.

2.2. Parametrization and rings of rank n. In terms of Theorem 2.1, it becomes easier to see what we might mean by "generalizations" of Gauss composition. Namely, we seek an algebraic group G and a representation V, defined over $\mathbb{Z}$, such that the set $G(\mathbb{Z}) \setminus V(\mathbb{Z})$ of integral orbits are in canonical bijection with interesting algebraic objects – such as rings of rank n, modules over these rings, and maps among them. In Gauss's case, the group G is SL_2 and V is the space of binary quadratic forms, and we have seen that the integral orbits parametrize oriented ideal classes (or oriented rank 1 modules) in quadratic rings. In general, we have the following question:

Question 2.2. For what pairs (G, V) does $G(\mathbb{Z}) \setminus V(\mathbb{Z})$ parametrize rings, modules, maps, etc.?

If other such pairs (G, V) do in fact exist, where do we go about looking for them? One thing to notice about the action of $\mathrm{GL}_2(\mathbb{C})$ on the vector space of binary quadratic forms over $\mathbb{C}$ is that there is essentially one (Zariski open) orbit – i.e., any binary quadratic form of nonzero discriminant can be taken to any other such form via an element of $\mathrm{GL}_2(\mathbb{C})$. It is also possible to see this from the point of view of Gauss composition: by "base change" the proof of Theorem 2.1 shows that nondegenerate orbits over $\mathbb{C}$ must be in one-to-one correspondence with quadratic rings over $\mathbb{C}$ – which must take the form $\mathbb{C} \oplus \mathbb{C}$ – and ideal classes over such rings – which also must take the form $\mathbb{C} \oplus \mathbb{C}$ (up to isomorphism). So over $\mathbb{C}$, there is essentially just one object of the form (S, I), namely $S = I = \mathbb{C} \oplus \mathbb{C}$.

By the same argument, if we are to get a parametrization result of a simple form like Gauss composition, where objects being parametrized are rings of rank n, ideal classes, etc. (so that there is only one such nondegenerate object over $\mathbb{C}$), then the pair (G, V) must also have the property that there is just one open orbit over $\mathbb{C}$. Such representations having just one open orbit over $\mathbb{C}$ have come up for numerous authors in various contexts, and they are known as "prehomogeneous vector spaces".

Definition 2.3. A *prehomogeneous vector space* is a pair (G, V) where G is an algebraic group and V is a rational vector space representation of G such that the action of $G(\mathbb{C})$ on $V(\mathbb{C})$ has just one Zariski open orbit.

In a monumental work, Sato and Kimura [33] gave a classification of all "reduced, irreducible" prehomogeneous vector spaces. Namely, they showed that there are essentially 36 of them! A few of these 36 are in fact infinite families. In another beautiful work, Wright and Yukie [41] studied these spaces over fields and found that the K-orbits for a field K frequently correspond to field extensions of K; for example, the nondegenerate $\mathbb{Q}$-orbits on the space of binary quadratic forms are naturally in bijection with quadratic extensions of $\mathbb{Q}$. So that gives us some hope, and obtaining the answer to Question 2.2 thus translates into the following goal:

Goal 2.4. Understand $G(\mathbb{Z}) \setminus V(\mathbb{Z})$ for prehomogeneous vector spaces (G, V).

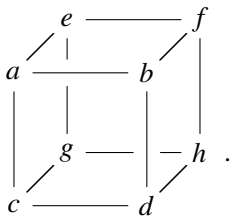
Of course, some of these spaces are quite large – thirty or more dimensions – so to just go in and analyze the integer orbits is somewhat daunting. Even Gauss's space,

which is only three-dimensional, is (as we have seen!) far from trivial. Gauss's own treatment of Gauss composition took numerous pages to describe.

To make further progress, we wish to have a different – and perhaps also simpler – perspective on Gauss composition that might lend itself more naturally to generalization to other spaces. Following [4], we give such a perspective in terms of $2 \times 2 \times 2$ cubes of integers. As we will see, the space of $2 \times 2 \times 2$ cubes not only gives an elementary description of Gauss composition, but also leads to composition laws and analogues of Theorem 2.1 for numerous other prehomogeneous vector spaces.

3. The story of the cube

Suppose we put integers on the corners of a cube:



(4)

Notice that any such cube A of integers may be sliced into two 2×2 matrices, and in essentially three different ways, corresponding to three different planes of symmetry of a cube. More precisely, the integer cube A given by (4) can be sliced into the following pairs of 2×2 matrices:

$$\begin{aligned} M_1 &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad N_1 = \begin{bmatrix} e & f \\ g & h \end{bmatrix} \\ M_2 &= \begin{bmatrix} a & c \\ e & g \end{bmatrix}, \quad N_2 = \begin{bmatrix} b & d \\ f & h \end{bmatrix} \\ M_3 &= \begin{bmatrix} a & e \\ b & f \end{bmatrix}, \quad N_3 = \begin{bmatrix} c & g \\ d & h \end{bmatrix}. \end{aligned} \tag{5}$$

Now for any such slicing of the cube A into a pair (M_i, N_i) of 2×2 matrices as in (5), we may construct a binary quadratic form $Q_i(x, y)$ as follows:

$$Q_i(x, y) = -\mathrm{Det}(M_i x + N_i y). \tag{6}$$

Thus any cube A of integers gives rise to three integral binary quadratic forms. A simple computation or elementary argument shows that the discriminants of the three quadratic forms Q_1, Q_2, and Q_3 are the same! And the punchline is:

Theorem 3.1. *If a cube A gives rise to three primitive binary quadratic forms Q_1, Q_2, Q_3 via* (4)–(6)*, then Q_1, Q_2, Q_3 have the same discriminant, and the product of these three forms is the identity in the group defined by Gauss composition.*

Conversely, if Q_1, Q_2, Q_3 are any three primitive binary quadratic forms of the same discriminant whose product is the identity under Gauss composition, then there exists a cube A yielding Q_1, Q_2, Q_3 via (4)–(6).

Thus the cube story gives a very simple and complete description of Gauss composition of binary quadratic forms. In fact, Theorem 3.1 can be used to *define* Gauss composition. The situation is reminiscent of the group law on a plane elliptic curve, where the most elementary way to define the group law is to declare that three points sum to zero if and only if they lie on a common line. In the same way, we may define Gauss composition by declaring that three primitive quadratic forms multiply to the identity if and only if they arise from a common cube. A proof of Theorem 3.1 may be found in [4, Appendix].

Theorem 3.1 is useful not only because it leads to Gauss composition, but also because it leads to various additional laws of composition. First and foremost, it leads to a law of composition on the cubes themselves!

3.1. Composition of cubes. Let us begin by rephrasing Theorem 3.1 as an orbit problem. First, we note that the space of cubes may be identified with the representation $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$ of the group $G = \mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z})$; this representation is a prehomogeneous vector space. The identification is made as follows: if we use $\langle v_1, v_2 \rangle$ to denote the standard basis of $\mathbb{Z}^2$, then the cube described by (4) is simply

$$\begin{aligned} &a\, v_1 \otimes v_1 \otimes v_1 + b\, v_1 \otimes v_2 \otimes v_1 + c\, v_2 \otimes v_1 \otimes v_1 + d\, v_2 \otimes v_2 \otimes v_1 \\ &+ e\, v_1 \otimes v_1 \otimes v_2 + f\, v_1 \otimes v_2 \otimes v_2 + g\, v_2 \otimes v_1 \otimes v_2 + h\, v_2 \otimes v_2 \otimes v_2 \end{aligned}$$

as an element of $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$. In terms of the cubical representation (4), the three factors of $\mathrm{SL}_2(\mathbb{Z})$ in G act by row operations, column operations, and the "other direction" operations respectively.

Theorem 3.1 may be viewed as describing the nondegenerate orbits of $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$ under the action of G in terms of triples of oriented ideal classes whose "product" is the identity class. To state this description more precisely, we need just two simple definitions. First, we call a triple (I_1, I_2, I_3) of oriented fractional ideals in $S \otimes \mathbb{Q}$ *balanced* if $I_1 I_2 I_3 \subseteq S$ and $\mathrm{N}(I_1)\mathrm{N}(I_2)\mathrm{N}(I_3) = 1$. Also, we define two balanced triples (I_1, I_2, I_3) and (I_1', I_2', I_3') of oriented ideals of S to be *equivalent* if $I_1 = \kappa_1 I_1'$, $I_2 = \kappa_2 I_2'$, and $I_3 = \kappa_3 I_3'$ for some invertible elements $\kappa_1, \kappa_2, \kappa_3 \in S \otimes \mathbb{Q}$. For example, if S is the full ring of integers in a quadratic field, then an equivalence class of balanced triples means simply a triple of oriented ideal classes whose product is the principal class.

Our Theorem 3.1 on cubes may then be stated as the solution to an orbit problem as follows:

Theorem 3.2. *There is a canonical bijection between the set of nondegenerate G-orbits on the space $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$ of $2 \times 2 \times 2$ integer cubes and the set of isomorphism classes of pairs $(S, (I_1, I_2, I_3))$, where S is a nondegenerate oriented quadratic ring and (I_1, I_2, I_3) is an equivalence class of balanced triples of oriented ideals of S.*

As with Theorem 3.2, we may consider those orbits that correspond solely to *invertible* oriented ideal classes. Let us say a cube A is *projective* if the three oriented ideal classes associated to A in Theorem 3.2 are invertible (i.e., if they are projective as modules). Equivalently, A is projective if the associated three binary quadratic forms Q_i are each primitive.

Let us define the *discriminant* $\mathrm{Disc}(A)$ of a cube A to be the discriminant of any one of the three binary quadratic forms Q_i arising from it. Then Theorem 3.2 is discriminant-preserving: under the bijection, the discriminant of a cube is equal to the discriminant of the corresponding quadratic ring.

We can now describe composition of cubes. It is most easily stated in terms of ideal classes. Recall that Gauss composition can be viewed as multiplication of oriented ideal classes in a fixed quadratic ring S:

$$(S, I) \circ (S, I') = (S, II').$$

When restricted to invertible ideal classes of a fixed quadratic ring $S = S(D)$ (i.e., primitive binary quadratic forms having a fixed discriminant D), this yields the oriented class group $\mathrm{Cl}^+(S(D))$.

Analogously, composition of cubes can be viewed as multiplication of equivalence classes of balanced triples of oriented ideals:

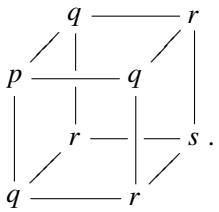
$$(S, (I_1, I_2, I_3)) \circ (S, (I_1', I_2', I_3')) = (S, (I_1I_1', I_2I_2', I_3I_3')).$$

When restricted to invertible ideal classes of a fixed quadratic ring (i.e., projective cubes having a fixed discriminant), this yields the group $\mathrm{Cl}^+(S) \times \mathrm{Cl}^+(S)$, since the last ideal class is determined by the first two. Thus Gauss composition yields $\mathrm{Cl}^+(S)$, while composition of cubes gives $\mathrm{Cl}^+(S) \times \mathrm{Cl}^+(S)$. A surprising consequence of this result is that the number of orbits of projective cubes having a given discriminant D is always a square number.

3.2. Composition of binary cubic forms. The law of composition of cubes now also leads to a number of further composition laws on various other spaces. First, let us consider the space of triply-symmetric cubes, which is equivalent to the space of binary cubic forms $px^3 + 3qx^2y + 3rxy^2 + sy^3$: indeed, just as one often expresses a binary quadratic form $px^2 + 2qxy + ry^2$ as the symmetric 2×2 matrix

$$\begin{bmatrix} p & q \\ q & r \end{bmatrix},$$

one may naturally express a binary cubic form $px^3 + 3qx^2y + 3rxy^2 + sy^3$ via the triply-symmetric $2 \times 2 \times 2$ matrix



(7)

If we use $\mathrm{Sym}^3\mathbb{Z}^2$ to denote the space of binary cubic forms with triplicate central coefficients, then the above association of $px^3 + 3qx^2y + 3rxy^2 + sy^3$ with the cube (7) corresponds to the natural inclusion

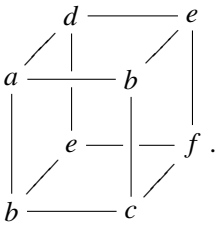
$$\iota\colon \mathrm{Sym}^3\mathbb{Z}^2 \to \mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$$

of the space of triply-symmetric cubes into the space of cubes. The space of binary cubic forms under the action of $\mathrm{SL}_2(\mathbb{Z})$ also yields a prehomogeneous vector space.

We call a binary cubic form $C(x, y) = px^3 + 3qx^2y + 3rxy^2 + sy^3$ *projective* if the corresponding triply-symmetric cube $\iota(C)$ given by (7) is projective. It turns out that the $\mathrm{SL}_2(\mathbb{Z})$-orbits on such binary cubic forms having a fixed discriminant D also then inherit a law of composition from the space of cubes, leading to a group structure when restricted to projective forms. It is not hard to guess what this group should be related to. Namely, projective triply-symmetric cubes correspond to a balanced triple of ideals (I, I, I) in $S(D)$, where the three ideals are in fact the same. Thus $I \cdot I \cdot I$ is the identity ideal class, implying that orbits of binary cubic forms essentially correspond to 3-torsion elements in the oriented class group $\mathrm{Cl}^+(S)$. (The precise 3-torsion group one obtains is discussed in [4].) Thus the symmetrization procedure allows us to isolate a certain arithmetic part of the class group.

An interesting consequence of this result is that the number of orbits of projective binary cubic forms having a given discriminant D is always a power of three!

3.3. Composition of pairs of binary quadratic forms. The group law on binary cubic forms of discriminant D was obtained by imposing a triple-symmetry condition on the group of $2 \times 2 \times 2$ cubes of discriminant D. Rather than imposing a threefold symmetry, one may instead impose only a twofold symmetry. This leads to cubes taking the form



(8)

That is, these cubes can be sliced (along a certain fixed plane) into two 2×2 symmetric matrices and therefore can naturally be viewed as a pair of binary quadratic forms $(ax^2 + 2bxy + cy^2, dx^2 + 2exy + fy^2)$.

If we use $\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^2$ to denote the space of pairs of integer-matrix binary quadratic forms, then the above association of $(ax^2+2bxy+cy^2, dx^2+2exy+fy^2)$ with the cube (8) corresponds to the natural inclusion map

$$\jmath : \mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^2 \to \mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2.$$

The lattice $\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^2$ under the action of $\mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z})$ again yields a prehomogeneous vector space.

As in the case of binary cubic forms, we call a pair of binary quadratic forms *projective* if the corresponding doubly-symmetric cube $\jmath(C)$ given by (8) is projective. Again, the projective pairs of binary quadratic forms having a fixed discriminant D inherit a group structure. Since such elements correspond to balanced triples of ideals (I_1, I_3, I_3) where the last two ideals are the same, one sees that the group thus obtained is again simply the group $\mathrm{Cl}^+(S(D))$ since I_3 determines I_1. That is, not only do binary quadratic forms of a fixed discriminant D give rise to the oriented class group of $S(D)$, but so do *pairs* of binary quadratic forms!

3.4. Further parametrization spaces for quadratic rings. The discussions above illustrate that once we have a law of composition on the space of cubes, then various other of its invariant and covariant spaces also inherit a law of composition; Gauss composition is indeed just one of these.

Symmetrization is one procedure that allows us to generate new prehomogenous vector spaces with composition; this was the subject of Sections 3.2 and 3.3. The determinant trick (6) to produce Gauss composition is another. There are several other operations too that play an important role, such as *skew-symmetrization*, *symplectization*, *hermitianization*, and *dualization*, and each procedure is found to have both invariant-theoretic and number-theoretic meaning, yielding numerous further analogues of Theorem 2.1 involving higher rank rings, higher rank modules, as well as noncommutative rings such as quaternion and octonion algebras. Further details may be found in [4] and [8].

4. Cubic analogues of Gauss composition

In the previous section, we discussed various generalizations of Gauss composition that were found to be closely related to the ideal class groups of quadratic rings. In this section, we show how similar ideas can be used to obtain genuine "cubic analogues" of Gauss composition, i.e., composition laws on appropriate spaces of forms so that the resulting groups are related to the class groups of *cubic rings*.

The fundamental object in our treatment of quadratic composition was the space of $2 \times 2 \times 2$ cubes of integers. It turns out that the fundamental object for cubic

composition is the space of $2\times3\times3$ boxes of integers, and yields exactly what is needed for a cubic analogue of Gauss's theory. The action of $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ on $2 \times 3 \times 3$ integer boxes is again a prehomogeneous vector space, and the orbits correspond in a natural way to cubic rings and ideal classes in those rings. Before the resulting cubic analogues of Gauss composition can be described, it is necessary first to understand how cubic rings are parametrized.

4.1. The parametrization of cubic rings. In Section 2.1, we saw that quadratic rings are parametrized up to isomorphism by their discriminants. This is not so for cubic rings; indeed, there may sometimes be several nonisomorphic cubic rings having the same discriminant. The correct object parametrizing cubic rings – i.e., rings free of rank 3 as $\mathbb{Z}$-modules – was first determined by Delone–Faddeev in their beautiful 1964 treatise on cubic irrationalites [21]. They showed that cubic rings are in bijective correspondence with $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of integral binary cubic forms $ax^3 + bx^2y + cxy^2 + dy^3$, as follows.

Given a binary cubic form $f(x, y) = ax^3+bx^2y+cxy^2+dy^3$ with $a, b, c, d \in \mathbb{Z}$, we associate to f the ring $R(f)$ having $\mathbb{Z}$-basis $\langle 1, \omega, \theta\rangle$ and multiplication table

$$\begin{aligned} \omega\theta &= -ad \\ \omega^2 &= -ac + b\omega - a\theta \\ \theta^2 &= -bd + d\omega - c\theta. \end{aligned} \tag{9}$$

One easily verifies that $\mathrm{GL}_2(\mathbb{Z})$-equivalent binary cubic forms then yield isomorphic rings, and conversely, that every isomorphism class of ring R can be represented in the form $R(f)$ for a unique binary cubic form f, up to such equivalence. Thus we may say that isomorphism classes of cubic rings are parametrized by $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of integral binary cubic forms. An easy calculation using (9) shows that the discriminant $\mathrm{Disc}(R(f))$ is equal to the discriminant $\mathrm{Disc}(f)$ of the binary cubic form f, where $\mathrm{Disc}(f) = b^2c^2 - 4ac^3 - 4b^3d + 18abcd - 27a^2d^2$ is the unique polynomial invariant for the action of $\mathrm{GL}_2(\mathbb{Z})$ on binary cubic forms. We thus obtain:

Theorem 4.1 ([21]). *There is a canonical bijection between the set of* $\mathrm{GL}_2(\mathbb{Z})$*-equivalence classes of integral binary cubic forms and the set of isomorphism classes of cubic rings, by the association*

$$f \leftrightarrow R(f).$$

Moreover, $\mathrm{Disc}(f) = \mathrm{Disc}(R(f))$.

We say a cubic ring is *nondegenerate* if it has nonzero discriminant (equivalently, if it is an order in an étale cubic algebra over $\mathbb{Q}$). Similarly, a binary cubic form is *nondegenerate* if it has nonzero discriminant (equivalently, if it has distinct roots in $\mathbb{P}^1(\mathbb{Q})$). The discriminant equality in Theorem 4.1 implies, in particular, that isomorphism classes of nondegenerate cubic rings correspond bijectively with equivalence classes of nondegenerate integral binary cubic forms.

4.2. Cubic composition and 2 × 3 × 3 boxes. Imitating Section 3.1, for a cubic ring R let us say a pair (I, I') of fractional R-ideals in $K = R \otimes \mathbb{Q}$ is *balanced* if $II' \subseteq R$ and $N(I)N(I') = 1$. Furthermore, we say two such balanced pairs (I_1, I_1') and (I_2, I_2') are *equivalent* if there exist invertible elements $\kappa, \kappa' \in K$ such that $I_1 = \kappa I_2$ and $I_1' = \kappa' I_2'$. For example, if R is the full ring of integers in a cubic field then an equivalence class of balanced pairs of ideals is simply a pair of ideal classes that are inverse to each other in the ideal class group.

The analogue of Theorem 3.2 in the theory of cubic composition states that $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$-classes of $2 \times 3 \times 3$ integer boxes correspond to equivalence classes of balanced pairs of ideals in cubic rings.

Theorem 4.2. *There is a canonical bijection between the set of nondegenerate* $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$*-orbits on the space* $\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3$ *and the set of isomorphism classes of pairs* $(R, (I, I'))$*, where* R *is a nondegenerate cubic ring and* (I, I') *is an equivalence class of balanced pairs of ideals of* R*.*

How does one recover the cubic ring R from a $2 \times 3 \times 3$ box of integers? A $2 \times 3 \times 3$ box may be viewed (by an appropriate slicing) as a pair (A, B) of 3×3 matrices. Then $f(x, y) = \mathrm{Det}(Ax - By)$ is a binary cubic form. The ring R is simply the cubic ring $R(f)$ associated to f via Theorem 4.1.

If we define the *discriminant* $\mathrm{Disc}(A, B)$ of (A, B) as $\mathrm{Disc}(\mathrm{Det}(Ax - By))$, then one shows again that this discriminant is the unique polynomial invariant for the action of $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ on $2 \times 3 \times 3$ boxes. By the method of recovering R from (A, B) above, we see again that the bijection of Theorem 4.2 preserves discriminants.

We may now describe composition of $2 \times 3 \times 3$ boxes. Given a binary cubic form f, let $(\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3)(f)$ denote the set of all elements $(A, B) \in \mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3$ such that $\mathrm{Det}(Ax - By) = f(x, y)$; all such elements correspond to the same cubic ring in Theorem 4.2. The group $\mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ is seen to act naturally on the set $(\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3)(f)$ via simultaneous row and column operations on A and B; this action evidently does not change the value of $\mathrm{Det}(Ax - By)$. Moreover, one finds that two elements of $(\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3)(f)$ yield equivalent balanced pairs of ideals in $R(f)$ if and only if they are equivalent under $\mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$.

As with the quadratic cases of composition in Section 3, composition of $2 \times 3 \times 3$ boxes having a fixed f can now be viewed as multiplication of equivalence classes of balanced pairs of ideals in the corresponding cubic ring $R = R(f)$:

$$(R, (I, I')) \circ (R, (J, J')) = (R, (IJ, I'J')).$$

When restricted to invertible ideal classes (i.e., *projective* $2 \times 3 \times 3$ boxes), this yields the ideal class group $\mathrm{Cl}(R)$ of R, since the second ideal class is determined by the first (as they are inverses to each other). Thus composition of $\mathrm{SL}_3(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$-equivalence classes of projective $2 \times 3 \times 3$ boxes yields the class groups of cubic rings, in complete analogy with Gauss composition in the quadratic case.

To summarize:

- In the case of binary quadratic forms, the unique SL_2-invariant is the discriminant D, which classifies orders in quadratic fields. The primitive classes having a fixed value of D form a group under a certain natural composition law. This group is naturally isomorphic to the narrow class group of the corresponding quadratic order.
- In the case of $2\times3\times3$ integer boxes, the unique $\mathrm{SL}_3\times\mathrm{SL}_3$-invariant is the binary cubic form f, which classifies orders in cubic fields. The projective classes having a fixed value of f form a group under a certain natural composition law. This group is naturally isomorphic to the ideal class group of the corresponding cubic order.

Thus the composition law on the space of $2 \times 3 \times 3$ integer cubes is really the cubic analogue of Gauss composition.

4.3. Cubic composition and pairs of ternary quadratic forms. Just as we were able to impose a symmetry condition on $2 \times 2 \times 2$ matrices to obtain information on the exponent 3-parts of class groups of quadratic rings, we can impose a symmetry condition on $2 \times 3 \times 3$ matrices to obtain information on the exponent 2-parts of class groups of cubic rings. The "symmetric" elements in $\mathbb{Z}^2 \otimes \mathbb{Z}^3 \otimes \mathbb{Z}^3$ are the elements of $\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3$, i.e., pairs (A, B) of symmetric 3×3 integer matrices, which may be viewed as pairs (A, B) of integral ternary quadratic forms. The action of $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ on pairs of ternary quadratic forms is again a prehomogeneous vector space.

The cubic form invariant f and the *discriminant* $\mathrm{Disc}(A, B)$ of (A, B) may be defined in the identical manner; we have $f(x, y) = \mathrm{Det}(Ax - By)$ and $\mathrm{Disc}(A, B) = \mathrm{Disc}(\mathrm{Det}(Ax - By))$. We say an element $(A, B) \in \mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3$ is *projective* if it is projective as a $2 \times 3 \times 3$ box.

As in the case of binary cubic forms and symmetric cubes (see Section 3.2), the space of pairs of ternary quadratic forms also inherits a law of composition from the space of $2 \times 3 \times 3$ boxes. Again, the restriction to symmetric classes isolates a certain arithmetic part of the class group. Namely, symmetric projective $2 \times 3 \times 3$ boxes yield pairs of the form $(R, (I, I))$ where the two ideals are in fact the same. Thus $I \cdot I$ is the identity ideal class of R, so we see that $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$-orbits of pairs of integer-matrix ternary quadratic forms essentially parametrize 2-torsion elements in the class groups of cubic rings (see [5] for further details).

This parametrization has several interesting consequences. For example, it implies that the number of equivalence classes of projective pairs (A, B) of ternary quadratic forms having a given binary cubic $\mathrm{Det}(Ax - By)$ is always a power of 2!

The parametrization also enables one to prove the first known case of the Cohen–Martinet heuristics for class groups, namely for the average size of the 2-torsion subgroup in the class groups of cubic fields. This average number of 2-torsion elements turns out to be 5/4 for real cubic fields and 3/2 for complex cubic fields. In particular,

this implies that at least 75% of totally real cubic fields, and at least 50% of complex cubic fields, have odd class number. Further details may be found in [9]. The case of narrow class groups can also be handled by generalizations of these arguments (to appear in future work).

5. The parametrization of quartic and quintic rings

The composition laws and results of the previous two sections depended heavily on the simple but beautiful parametrizations of quadratic and cubic rings given by (2) and (9) respectively. Namely, we saw that quadratic rings are parametrized by integers congruent to 0 or 1 (mod 4), while cubic rings are parametrized by $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of binary cubic forms.

It has been a long-time open question to determine whether analogous parametrizations exist for rings of rank 4. The ideas of the previous sections, together with a theory of *resolvent rings*, lead to a parametrization of quartic rings that is just as complete as in the quadratic and cubic cases. These "resolvent rings" are so named because they form natural integral models of the resolvent fields occurring in the classical literature; see [6] for further details.

This perspective leads one to show that the analogous objects parametrizing quartic rings are essentially *pairs of integer-valued ternary quadratic forms*, up to integer equivalence. To make a precise statement, let $(\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3)^*$ denote the space of pairs of ternary quadratic forms having integer coefficients. Then we have:

Theorem 5.1. *There is a canonical bijection between the set of* $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$*-orbits on the space* $(\mathbb{Z}^2 \otimes \mathrm{Sym}^2\mathbb{Z}^3)^*$ *of pairs of integer-valued ternary quadratic forms and the set of isomorphism classes of pairs* (Q, R)*, where* Q *is a quartic ring and* R *is a cubic resolvent ring of* Q*.*

A cubic resolvent ring of a quartic ring Q is a cubic ring R equipped with a certain natural quadratic mapping $Q \to R$. It turns out that all quartic rings have at least one cubic resolvent ring; moreover, for "most" quartic rings (e.g., for maximal quartic rings) this cubic resolvent ring is in fact unique. Thus every quartic ring arises in Theorem 5.1, and the theorem yields a bijection on the quartic rings of primary interest to algebraic number theorists, namely the maximal orders in quartic fields.

The theory of resolvent rings used in [6] to prove Theorem 5.1 makes heavy use of many of the formulae arising in the solution to the quartic equation. The same ideas also yield a purely ring-theoretic interpretation of the Delone–Faddeev parametrization of cubic rings, using formulae arising in the classical solution to the cubic equation.

Since the quintic equation is known to be "unsolvable", it may then seem that such parametrization methods could not extend beyond the quartic. However, there has been a lot of literature on the quintic equation, and some of the formulae that arise

in these works – although they fail to "solve" the quintic equation – can nevertheless be adapted to develop a completely analogous theory for parametrizing quintic rings! It turns out that quintic rings are essentially parametrized by *quadruples of quinary alternating bilinear forms*, i.e., quadruples of 5×5 skew-symmetric integer matrices.

Let $\mathbb{Z}^4 \otimes \wedge^2\mathbb{Z}^5$ denote the space of quadruples of 5×5 skew-symmetric integer matrices. Then the parametrization result for quintic rings is as follows.

Theorem 5.2. *There is a canonical bijection between the set of* $\mathrm{GL}_4(\mathbb{Z}) \times \mathrm{SL}_5(\mathbb{Z})$*-orbits on the space* $\mathbb{Z}^4 \otimes \wedge^2\mathbb{Z}^5$ *of quadruples of* 5×5 *skew-symmetric integer matrices and the set of isomorphism classes of pairs* (R, S)*, where* R *is a quintic ring and* S *is a sextic resolvent ring of* R*.*

A *sextic resolvent ring* of a quintic ring R is a sextic ring S equipped with a certain natural mapping $R \to \wedge^2 S$ which seems to have been missed in the classical literature on the quintic equation. The notion of sextic resolvent ring yields a natural integral model for the sextic resolvent fields studied by Cayley and Klein. As in the quartic case, one finds that all quintic rings have a sextic resolvent, and maximal quintic rings have exactly one sextic resolvent ring. Thus Theorem 5.2 yields a bijection on maximal orders in quintic fields!

These parametrization results have an important application to determining the density of discriminants of number fields of degree less than or equal to five, which we discuss in the next section.

Because of the classification of prehomogeneous vector spaces, one can show that parametrizations of the *same type* cannot exist for rings of rank $n > 5$. This is in agreement with the classification of group stabilizers by Sato–Kimura [33], and with the classification of orbits over fields by Wright–Yukie [41]. Thus parametrizations of this type end with the quintic.

6. Counting number fields of low degree

Number fields – i.e., field extensions of the rational numbers of finite degree – are perhaps the most fundamental objects in algebraic number theory, yet very little is known about their distribution with respect to basic invariants.

The most fundamental numerical invariant of a degree n number field K is its *discriminant* $\mathrm{Disc}(K)$. The quantity $\mathrm{Disc}(K)$ is defined as $\mathrm{Disc}(O_K)$, where O_K denotes the unique maximal ring of rank n contained in K (equivalently, O_K is the ring of algebraic integers in K). A fundamental theorem of Minkowski states that, up to isomorphism, there can be only finitely many number fields having any given discriminant D. The question thus arises: how many? The number of number fields of discriminant D fluctuates with D in a seemingly random manner, so that obtaining an exact answer would be rather unwieldy. Nevertheless, it is still natural to ask whether one can understand the answer on average. That is, how many number fields do we expect having discriminant D?

It is natural to refine the latter question by considering each degree and each associated Galois group separately. For the remainder of this section, we fix the degree to be n and consider the degree n number fields whose Galois closures have Galois group S_n, which is in some sense the "generic" case. We now consider successive cases of n, starting with the simplest case, namely

***n* = 1.** There is only one degree 1 number field, namely the field $\mathbb{Q}$ of rational numbers, and its discriminant is 1. Thus we expect zero degree 1 number fields per discriminant as the discriminant tends to infinity. □

***n* = 2.** The case $n = 2$ is also not difficult to handle. Recall that, for each nonsquare discriminant D, there is a unique quadratic order having discriminant D. Maximal orders correspond to discriminants that are not square multiples of other discriminants, so that maximality essentially amounts to a squarefree condition on D.[5] It is known that the probability that a number is squarefree is $6/\pi^2$; it follows that we expect about $6/\pi^2 \approx .607\ldots$ quadratic fields per discriminant. □

In the 1960s, the cases $n = 1$ and $n = 2$ apparently provided enough evidence for the following bold folk conjecture to come into existence. The origin of this conjecture seems to be unknown.

Conjecture 6.1. Let $N_n(X)$ denote the number of S_n-number fields of degree n having absolute discriminant at most X. Then

$$c_n = \lim_{X\to\infty} \frac{N_n(X)}{X}$$

exists, and is positive for $n \geq 2$.

That is, we expect about c_n S_n-number fields of degree n per discriminant, where c_n is some positive constant when $n > 1$. One question that immediately arose from the circulation of this conjecture was: what should the value of c_n be? Evidently $c_1 = 0$ and $c_2 = 6/\pi^2$, but no general formula for the value of c_n was known.

***n* = 3.** Some further data as to the nature of c_n was provided in the seminal 1970 work of Davenport and Heilbronn [20], who explicitly determined the value of c_3. A key ingredient in their work was the parametrization of cubic orders by $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of binary cubic forms (see Section 4.1).

To count the number of $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of binary cubic forms having absolute discriminant less than X, Davenport constructed a fundamental domain $\mathcal{F}$ for the action of $\mathrm{GL}_2(\mathbb{Z})$ on the four-dimensional real vector space V of binary cubic

[5] The precise condition is that the number be squarefree and 1 (mod 4) or be four times an integer that is 2 or 3 (mod 4). So it is not quite a squarefree condition at 2. Nevertheless, the density of such numbers is still $6/\pi^2$! This is *not* a coincidence, but is part of a general phenomenon occurring in all degrees and at all primes dividing the degree, which may be explained via certain "mass formulae" arising in work of Serre. More details may be found in [11].

forms over $\mathbb{R}$. The number of cubic orders having absolute discriminant at most X is then simply the number of integer points in the region $\mathcal{F}_X$, where

$$\mathcal{F}_X = \mathcal{F} \cap \{v \in V : |\mathrm{Disc}(v)| \leq X\}.$$

This region is seen to have finite volume, namely $(\pi^2/18)X$.

Now given any region $\mathcal{R}$ in n-dimensional Euclidean space, it is very natural to approximate the number of integer lattice points in $\mathcal{R}$ by the Euclidean volume $\mathrm{Vol}(\mathcal{R})$. Such an approximation will be particularly good if the region is compact and somewhat "round-looking" in the sense that its boundaries are smooth, and there are no serious "spikes" or "tentacles" jutting out of the region.

However, if the region is noncompact or it possesses thin, long tentacles or spikes, then all bets are off. For example, one may have a region with a tentacle thinning as it runs off to infinity, which has finite volume yet contains an infinite number of lattice points. Or one could have a region with one infinitely long tentacle, which has arbitrarily large (finite or infinite) volume yet contains *no* lattice points! It is easy to draw pictures even in two-dimensional space that illustrate each of these unruly scenarios. For such "bad" regions, there may be little correlation between the volume and the number of lattice points lying within.

Let us consider again Davenport's domain $\mathcal{F}_X$. If this subset of V were compact and round, we could then conclude that the number of lattice points within is essentially $(\pi^2/18)X$. However, the region $\mathcal{F}_X$ is not compact. Although we do not attempt to draw this region here – as it is four-dimensional – it is nevertheless easy to visualize roughly what this region looks like. Namely, $\mathcal{F}_X$ is relatively round-looking, but there is a single problematic tentacle going off to infinity (arising from the fact that $\mathrm{SL}_2(\mathbb{Z}) \setminus \mathrm{SL}_2(\mathbb{R})$ is noncompact). Thus, to make any conclusions regarding the number of lattice points in $\mathcal{F}_X$, it is necessary to deal with this tentacle.

What Davenport shows is that although this tentacle (or *cusp*) contains a very large number of lattice points, nearly all of these lattice points are *reducible* cubic forms; i.e., they correspond to cubic rings sitting not in a cubic field, but in the direct sum of $\mathbb{Q}$ and a quadratic field. Only a negligible number of irreducible points are found to lie in the cusp. Meanwhile, the lattice points in the main body of the region and away from the cusps correspond almost entirely to irreducible points, i.e., orders in cubic fields; only a negligible number of points in this main body are reducible.[6]

It follows that, as $X \to \infty$, one may approximate the number of irreducible points in $\mathcal{F}_X$ by the volume of the main body of the region. As the above cusps are found to have negligible volume, we conclude that the number of irreducible points in $\mathcal{F}_X$ is $(\pi^2/18)X$, where the error is $o(X)$. We therefore obtain

Theorem 6.2. *The number of cubic orders* (*in cubic fields*) *having absolute discriminant at most X is asymptotic to $(\pi^2/18)X$ as $X \to \infty$.*

[6]Here, by negligible, we mean "$o(X)$".

To pass from such cubic orders to maximal cubic orders (and thus to cubic fields) requires a delicate sieve, which was carried out in the remarkable work of Davenport–Heilbronn. The result of this sieve is:

Theorem 6.3 (Davenport–Heilbronn [20]). *The number of cubic fields having absolute discriminant at most X is asymptotic to $(1/3\zeta(3))X$ as $X \to \infty$, where $\zeta(s)$ denotes the Riemann zeta function.*

Thus Davenport and Heilbronn showed, in sum, that $c_3 = \frac{1}{3\zeta(3)} \approx .277\ldots$. That is, we expect approximately .277 cubic fields per discriminant. □

It has been a long-time open problem to extend Davenport–Heilbronn's theorem to $n = 4$, i.e., to gain an understanding of quartic number fields in the same way. Having now obtained a parametrization of quartic orders, it is natural to try and proceed in a manner analogous to Davenport–Heilbronn.

***n* = 4.** As discussed in Section 5, quartic orders are parametrized by pairs of integer-coefficient ternary quadratic forms, modulo the action of the group $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$. In analogy with Davenport–Heilbronn's work, we construct a fundamental domain $\mathcal{F}$ for the action of $\mathrm{GL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ on the space V of pairs of real-coefficient ternary quadratic forms. Then $\mathcal{F}$ is a certain region in the 12-dimensional real vector space V, and quartic rings are seen to correspond to lattice points inside the fundamental domain $\mathcal{F}$.

In order to understand the number of quartic orders (and eventually quartic fields) having absolute discriminant at most X, we therefore wish to count the number of integer points inside the region $\mathcal{F}_X$, where the region $\mathcal{F}_X$ is as usual defined by $\mathcal{F} \cap \{v \in V : |\mathrm{Disc}(v)| \leq X\}$. However, the region $\mathcal{F}_X$ is not so simple; indeed, the geometry of this fundamental domain is significantly more complicated than the analogous region considered by Davenport and Heilbronn. For one thing, the dimension is now twelve instead of four! Moreover, there are now three major cusps or tentacles rather than one, and the cross sections of these cusps lie in several dimensions. If these cusps did not exist, and $\mathcal{F}_X$ were compact, it would be an easy matter to estimate the number of integer points in $\mathcal{F}_X$.

It takes quite a bit of hard work to deal with the cusps, but in the end, what happens with these cusps is quite beautiful. All three cusps contain *many* points (i.e., at least on the order of X in number). However, essentially all the lattice points in the first cusp are found to be "reducible": they correspond to quartic rings that lie in the direct sum of two quadratic fields instead of a single quartic field. The second cusp also consists almost entirely of "reducible" points! These points correspond to orders lying in the direct sum of $\mathbb{Q}$ and a cubic field (or some other étale cubic algebra) rather than a quartic field. In the third cusp, another very interesting phenomenon occurs, namely the lattice points inside almost entirely correspond to orders in quartic fields whose Galois closure has Galois group D_4 (the dihedral group of order 8) rather than S_4! Meanwhile, the main body of the region away from the cusps is shown to consist

almost entirely (i.e., up to $o(X)$) of the lattice points that correspond to orders in S_4-quartic fields.[7]

As a result, to count orders in S_4-quartic fields (i.e., "S_4-quartic orders"), one may simply count lattice points in the region $\mathcal{F}_X$ with its tentacles cut off. This region is then compact, and is sufficiently round for one to deduce that the number of lattice points inside this region is essentially its volume, which is computed to be $(5\zeta(2)^2\zeta(3)/24)X$. It follows (in conjunction with Theorem 5.1) that the number of pairs (Q, R), where Q is an S_4-quartic order of discriminant at most X and R is a cubic resolvent ring of Q, is asymptotic to $(5\zeta(2)^2\zeta(3)/24)X$ as $X \to \infty$.

Finally, one shows that counting quartic rings just once each – i.e., without weighting by the number of cubic resolvents – affects this final answer simply by a factor of $\zeta(5)$. We obtain:

Theorem 6.4. *The number of S_4-quartic orders having absolute discriminant at most X is asymptotic to $\frac{5\zeta(2)^2\zeta(3)}{24\,\zeta(5)}\,X$ as $X \to \infty$.*

To count only the maximal orders in S_4-quartic fields again requires a fairly delicate sieve. The end result of this sieve is:

Theorem 6.5. *The number of S_4-quartic fields having absolute discriminant at most X is asymptotic to $\frac{5}{24}\prod_p(1 + p^{-2} - p^{-3} - p^{-4}) \cdot X$ as $X \to \infty$.*

Thus $c_4 = \frac{5}{24}\prod_p(1 + p^{-2} - p^{-3} - p^{-4}) \approx .253\ldots$; that is, we expect about .253 S_4-quartic fields per discriminant. □

Theorem 6.5 has a number of interesting consequences. First, it is related to the proof of the case of the Cohen–Lenstra–Martinet class group heuristics mentioned at the end of Section 4.3. Second, in conjunction with the work of Baily [1] and Cohen–Diaz–Olivier [13] on D_4-fields, Theorem 6.5 implies that when all quartic fields are ordered by the size of their discriminants, a positive proportion of them *do not* have Galois group S_4! In fact, "only" about 90.644% have Galois group S_4, while the remaining correspond to the Galois group D_4 (0% have any of the other possible Galois groups). This is interesting because it is in direct opposition to the situation for polynomials, where Hilbert's irreducibility theorem implies that if integer polynomials of degree n are ordered by the size of their coefficients, then 100% will have Galois group S_n.

***n* = 5.** Last but not least, the parametrization results described in the previous section also allow one to asymptotically determine the number of quintic fields of bounded discriminant. This represents the first instance where one can count *unsolvable* extensions.

[7] In practice, to simplify the details, we perform all these computations using not one but several fundamental domains $\mathcal{F}_X \subset \mathbb{R}^{12}$, but the spirit of the argument remains unchanged. The details of this "averaging" method may be found in [9] and [10].

We have shown that quintic rings correspond to the $\mathrm{GL}_4(\mathbb{Z}) \times \mathrm{SL}_5(\mathbb{Z})$-orbits on quadruples of 5×5 skew-symmetric integer matrices. Following the cases $n = 3$ and $n = 4$, we begin by constructing a fundamental domain for the action of $\mathrm{GL}_4(\mathbb{Z}) \times \mathrm{SL}_5(\mathbb{Z})$ on the corresponding forty-dimensional real vector space V. We wish to understand the number of integer points in $\mathcal{F}_X$, where as before $\mathcal{F}_X = \mathcal{F} \cap \{v \in V : |\mathrm{Disc}(v)| \leq X\}$. This turns out to be significantly more difficult than the corresponding problem in the cubic and quartic cases. Besides being highly non-compact, the forty-dimensional fundamental domain $\mathcal{F}_X$ has an intriguingly complex system of numerous high-dimensional tentacles and cusps!

But in the end these cusps too exhibit a surprisingly beautiful structure, and can be handled in much the same way as in the cubic and quartic cases. Namely, we cut up this system of cusps into a finite number (approximately 160) of sub-cusps, all of which run off to infinity and thus present a problematic tentacle-type scenario. In each one of these 160 sub-cusps, one shows either that there are a negligible number of points within, *or* that essentially all points in that tentacle are reducible in a certain way. In this aspect, these cusps are very similar to those occurring in the cases $n = 3$ and $n = 4$ – the difference being only that there are *many* more of them this time!

Lastly, one shows that 100% of the points in the main body of the region correspond to orders in S_5-quintic fields. By computing the volume of this main body, and then sieving down to the maximal quintic orders (for details on these tasks, see [10]), one finally obtains the following theorem.

Theorem 6.6. *The number of quintic fields having absolute discriminant at most X is asymptotic to $\frac{13}{120} \prod_p (1 + p^{-2} - p^{-4} - p^{-5}) \cdot X$ as $X \to \infty$.*

Therefore $c_5 = \frac{13}{120} \prod_p (1 + p^{-2} - p^{-4} - p^{-5}) \approx .149\ldots$, and so there are about .149 quintic fields per discriminant on average. □

***n* ≥ 6?** Given the success in the cases $n \leq 5$, the question is only too tempting: what happens for $n \geq 6$? We put forth the following conjecture:

Conjecture 6.7. We have

$$c_n = \frac{r_2(S_n)}{2 \cdot n!} \prod_p \Big(\sum_{k=0}^{n} \frac{q(k, n-k) - q(k-1, n-k+1)}{p^k} \Big) \tag{10}$$

where $q(i, j)$ denotes the number of partitions of i into at most j parts, and $r_2(S_n)$ denotes the number of 2-torsion elements in S_n.

That is, we conjecture that the number of S_n-number fields per discriminant will be c_n on average, where c_n is given as in (10).

Conjecture 6.7 was obtained by combining global heuristics with new mass formulae for étale extensions of local fields inspired by work of Serre [36]. It is readily checked that Conjecture 6.7 agrees with the values of c_n now proven for $n = 1$ through 5. For further details on this conjecture and the related mass formulae, see [11]. The proofs of the conjecture for $n = 3$, 4, and 5 may be found in [20], [9], and [10] respectively.

7. Related and future work

The composition and parametrization laws described in Sections 3–5 all turn out to be closely related to certain exceptional Lie groups. More precisely, let E be an exceptional Lie group and let P be a maximal parabolic of E. If we write $E = GU$, where G is the Levi factor and U is the unipotent radical at P, then the group G acts naturally (by conjugation) on the abelianized unipotent radical $V = U/[U, U]$. For appropriate choices of E and P, we find that we obtain precisely the prehomogeneous vector spaces (G, V) underlying the composition laws and parametrizations described in Sections 3–5. For example, the first case we considered in Section 3 was the space of $2 \times 2 \times 2$ cubes, and this representation of $\mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z})$ arises in this way when E is the exceptional Lie group of type D_4 and P is the Heisenberg parabolic.

This remarkable connection with Lie groups in fact appears to run much further – see [4, §4] and [5, §4] – and needs exploration, perhaps in connection with automorphic forms on these groups in the sense of Gan–Gross–Savin [26] and in the subsequent work of Lucianovic [31] and Weissman [38]. The reduction-theoretic aspect of some of the exceptional representations that arise in this way and their relation to noncommutative rings has been the subject of study in the recent work of Krutelevich.

The parametrization results described in Sections 3–5 for commutative rings extend to a large extent also to many noncommutative rings such as quaternions, octonions, and higher rank division algebras (see [8]). Many of these parametrizations of noncommutative rings and modules were discovered by applying certain number-theoretic operations (mentioned in Section 3.4) to parametrizations involving quadratic and cubic rings, indicating that there is a great deal of number theory lurking behind noncommutative – and even nonassociative! – rings such as the octonions. These number-theoretic connections beg for further investigation.

We note that the spaces underlying these various parametrizations also come equipped with a theory of zeta functions. Zeta functions associated to prehomogeneous vector spaces were first introduced by Sato and Shintani [34], and were further developed by Datskovsky, Wright, Yukie, and others. In particular, Datskovsky and Wright [16] used such zeta functions to give an alternative proof of Davenport–Heilbronn's theorem, which applies over an arbitrary number field or function field. The more difficult quartic analogue of their work was initiated by Yukie [42], and could eventually lead to an alternative method for counting quartic fields. The problem of understanding the relationship between the various parametrizations discussed here and the associated zeta functions is intriguing and deserves further investigation, both in the commutative and noncommutative cases.

Regarding commutative rings, the problem of finding parametrizations for rings of rank > 5 is a very interesting one. Although we have already noted that commutative rings cannot be parametrized by prehomogeneous vector spaces beyond the quintic case, there may be other ways to accomplish the task, such as through the study of

integer points on certain special varieties. This is a central problem in the theory, and of importance not only algebraically but also with respect to understanding the distribution of algebraic number rings and fields of higher degree.

As to our Conjecture 6.7 on counting S_n-number fields having fixed degree n and absolute discriminant less than X, even the correct order of growth (i.e., $O(X)$) is not known. The best general bounds known for $n \geq 6$ are due to Ellenberg and Venkatesh [24], who prove a bound of $O(X^{n^\varepsilon})$. Conjectures for the density of discriminants of degree n number fields having a specified Galois group G (yielding the expected orders of growth but not the constants) have been suggested by Malle [32]. These conjectures have been proven for many specific cases of G, including S_n for $n \leq 5$ (see Section 6), abelian groups (Wright [40]), D_4 (Cohen–Diaz–Olivier [13]), and certain nilpotent groups (Klüners–Malle [30]). The constants $c(G)$ in these conjectures for $G \neq S_n$ are unknown in general, even conjecturally, although there has been some recent progress. If the case $G = S_n$ is any indication, the constants $c(G)$ likely contain a great deal of arithmetic information.

One important ingredient in the case $G = S_n$ in determining the corresponding constants $c_n = c(S_n)$ was the development of mass formulae that count étale extensions of local fields by appropriate weights (see [11]). How these mass formulae change with G is an intriguing question, and several interesting cases and families of finite groups G have been treated by Kedlaya [29] and more recently by Wood. The manner in which these various local mass formulae glue together globally to give the global constants $c(G)$ is still an open question.

The counting arguments described in Section 6 can be taken much further, leading e.g. to further information on the distribution of class numbers, narrow class numbers, units, and regulators of cubic rings and fields. They can also be used to obtain information on the discriminant density of noncommutative rings such as quaternion and octonion rings, and modules over these rings. Finally, results analogous to those described in this survey can be obtained for ring and field extensions not just over $\mathbb{Z}$ and $\mathbb{Q}$ but over more general base rings. These directions too must be investigated, and we hope to treat them further in future work.

References

[1] Baily, A. M., On the density of discriminants of quartic fields. *J. Reine Angew. Math.* **315** (1980), 190–210.

[2] Belabas, K., Bhargava, M., Pomerance, C., Error terms for the Davenport-Heilbronn theorems. Preprint.

[3] Bhargava, M., Higher Composition Laws. Ph.D. Thesis, Princeton University, 2001.

[4] Bhargava, M., Higher composition laws I: A new view on Gauss composition. and quadratic generalizations. *Ann. of Math.* **159** (1) (2004), 217–250.

[5] Bhargava, M., Higher composition laws II: On cubic analogues of Gauss composition. *Ann. of Math.* **159** (2) (2004), 865–886.

[6] Bhargava, M., Higher composition laws III: The parametrization of quartic rings. *Ann. of Math.* **159** (3) (2004), 1329–1360.

[7] Bhargava, M., Higher composition laws IV: The parametrization of quintic rings. *Ann. of Math.*, to appear.

[8] Bhargava, M., Higher composition laws V: The parametrization of quaternionic and octonionic rings and modules. In preparation.

[9] Bhargava, M., The density of discriminants of quartic rings and fields. *Ann. of Math.* **162** (2005), 1031–1063.

[10] Bhargava, M., The density of discriminants of quintic rings and fields. *Ann. of Math.*, to appear.

[11] Bhargava, M., Mass formulae for local extensions and conjectures on the density of number field discriminants. Preprint.

[12] Brahmagupta, *Brahma-sphuṭa-siddhānta*. 628.

[13] Cohen, H., Diaz y Diaz, F., Olivier, M., Counting discriminants of number fields of degree up to four. In *Algorithmic Number Theory* (Leiden, 2000), Lecture Notes in Comput. Sci. 1838, Springer-Verlag, New York 2000, 269–283.

[14] Cohen, H., Martinet J., Étude heuristique des groupes de classes des corps de nombres. *J. Reine Angew. Math.* **404** (1990), 39–76.

[15] Cohen, H., Martinet J., Heuristics on class groups: some good primes are not too good. *Math. Comp.* **63** (1994), 329–334.

[16] Datskovsky, B., Wright, D. J., The adelic zeta function associated to the space of binary cubic forms II. Local theory. *J. Reine Angew. Math.* **367** (1986), 27–75.

[17] Davenport, H., On a principle of Lipschitz. *J. London Math. Soc.* **26** (1951), 179–183.

[18] Davenport, H., On the class-number of binary cubic forms I and II. *J. London Math. Soc.* **26** (1951), 183–198.

[19] Davenport, H., Corrigendum: "On a principle of Lipschitz". *J. London Math. Soc.* **39** (1964), 580.

[20] Davenport, H., Heilbronn, H., On the density of discriminants of cubic fields II. *Proc. Roy. Soc. London Ser. A* **322** (1551) (1971), 405–420.

[21] Delone, B. N., Faddeev, D. K., *The theory of irrationalities of the third degree*. Transl. Math. Monographs 10, Amer. Math. Soc., Providence, R.I., 1964

[22] Dirichlet, P. G. L., *Zahlentheorie*. 4th. edition, Vieweg, Braunschweig 1894.

[23] Eisenstein, G., Théorèmes sur les formes cubiques et solution d'une équation du quatrième degré indéterminées. *J. Reine Angew. Math.* **27** (1844), 75–79.

[24] Ellenberg, J., Venkatesh, A., The number of extensions of a number field with fixed degree and bounded discriminant. *Ann. of Math.* **163** (2) (2006), 723–741.

[25] Ennola, V., Turunen, R., On totally real cubic fields. *Math. Comp.* **44** (1985), 495–518.

[26] Gan, W.-T., Gross, B. H., Savin, G., Fourier coefficients of modular forms on G_2. *Duke Math. J.* **115** (1) (2002), 105–169.

[27] Gauss, C. F., *Disquisitiones Arithmeticae*. Leipzig, 1801.

[28] Hilbert, D., *Theory of Algebraic Invariants*. Engl. trans. by R. C. Laubenbacher, Cambridge University Press, Cambridge 1993.

[29] Kedlaya, K. S., Mass formulas for local Galois representations (after Serre, Bhargava). Preprint.

[30] Klüners, J., Malle, G., Counting nilpotent Galois extensions. *J. Reine Angew. Math.* **572** (2004), 1–26.

[31] Lucianovic, M., Quaternion Rings, Ternary Quadratic Forms, and Fourier Coefficients of Modular Forms on PGSp(6). Ph.D. Thesis, Harvard University, 2003.

[32] Malle, G., On the distribution of Galois groups. *J. Number Theory* **92** (2002), 315–329.

[33] Sato, M., Kimura, T., A classification of irreducible prehomogeneous vector spaces and their relative invariants. *Nagoya Math. J.* **65** (1977), 1–155.

[34] Sato, M., Shintani, T., On zeta functions associated with prehomogeneous vector spaces. *Ann. of Math.* (2) **100** (1974), 131–170.

[35] Serre, J-P., Modules projectifs et espaces fibrés à fibre vectorielle. *Séminaire Dubreil-Pisot* 1957/58, no. 23.

[36] Serre, J-P., Une "formule de masse" pour les extensions totalement ramifiées de degré donné d'un corps local. *C. R. Acad. Sci. Paris Sér. A-B* **286** (22) (1978), A1031–A1036.

[37] Vinberg, E. B., On the classification of the nilpotent elements of graded Lie algebras. *Soviet Math. Dokl.* **16** (1975), 1517–1520.

[38] Weissman, M., D_4 Modular Forms. *Amer. J. Math*, to appear.

[39] Wong, S., Automorphic forms on GL(2) and the rank of class groups. *J. Reine Angew. Math.* **515** (1999), 125–153.

[40] Wright, D. J., Distribution of discriminants of abelian extensions. *Proc. London Math. Soc.* (3) **58** (1989), 17–50.

[41] Wright, D. J., Yukie, A., Prehomogeneous vector spaces and field extensions. *Invent. Math.* **110** (1992), 283–314.

[42] Yukie, A., *Shintani Zeta Functions*. London Math. Soc. Lecture Note Ser. 183, Cambridge University Press, Cambridge 1993.

[43] Yukie, A., Density theorems for prehomogeneous vector spaces. Preprint.

Department of Mathematics, Princeton University, Princeton, NJ 08544, U.S.A.
E-mail: bhargava@math.princeton.edu

Hecke orbits as Shimura varieties in positive characteristic

Ching-Li Chai*

Abstract. Let p be a prime number, and let $\mathcal{M}$ be a modular variety of PEL type over $\overline{\mathbb{F}}_p$ which classifies abelian varieties in characteristic p with extra symmetries of a fixed PEL type. Consider the p-divisible group with extra symmetries consisting of all p-power torsions of the universal abelian scheme over $\mathcal{M}$. The locus in $\mathcal{M}$ corresponding to a fixed isomorphism type of a p-divisible group with extra symmetry is called a *leaf* by F. Oort. Each leaf is a smooth locally closed subvariety of the modular variety $\mathcal{M}$ which is stable under all prime-to-p Hecke correspondences on $\mathcal{M}$. Oort conjectured that every Hecke orbit is dense in the leaf containing it. Tools fashioned for this conjecture include (a) rigidity, (b) global monodromy, and (c) canonical coordinates. The theory of canonical coordinates generalizes the classical Serre–Tate coordinates; it asserts that locally at the level of jet-spaces, every leaf is built up from p-divisible formal groups through a finite family of fibrations in a canonical way. The Hecke orbit conjecture is affirmed when $\mathcal{M}$ is a Siegel modular variety classifying principally polarized abelian varieties of a fixed dimension, and also when $\mathcal{M}$ is a Hilbert modular variety classifying abelian varieties with real multiplications. The proof of the Siegel case, joint with F. Oort, uses the irreducibility of non-supersingular leaves in Hilbert modular varieties due to C.-F. Yu. That proof relies heavily on a special property of Siegel modular varieties: The set of $\overline{\mathbb{F}}_p$-rational points of a Siegel modular variety $\mathcal{A}_{g,n}$ is filled up by $\overline{\mathbb{F}}_p$-rational points of Hilbert modular varieties contained in $\mathcal{A}_{g,n}$. Possible directions for further progress include Tate-linear subvarieties and p-adic monodromy. The title of this article suggests that each leaf deserves to be viewed as a Shimura variety in characteristic p in its own right.

Mathematics Subject Classification (2000). Primary 14K10; Secondary 11G10.

Keywords. Shimura variety, Hecke correspondence, moduli, leaf, Barsotti–Tate group, abelian variety, monodromy.

1. Introduction

Let p be prime number and let $k \supset \mathbb{F}_p$ be an algebraically closed field fixed throughout this article; the field k will serve as the base field of modular varieties. The reader may want to take $k = \overline{\mathbb{F}}_p$.

We are interested in Hecke symmetries on the reduction to k of a Shimura variety. Because the theory of integral models of Shimura varieties are not fully developed yet, we will restrict to a small class of Shimura varieties, call modular varieties of PEL

*Partially supported by NSF grant DMS04-00482.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

type. Such a modular variety classifies abelian varieties with prescribed polarization and endomorphisms of a fixed type.

We will further restrict our attention to the prime-to-p Hecke symmetries. Since these symmetries come from finite étale isogeny correspondences for the universal abelian scheme over a modular variety $\mathcal{M}$, they preserve all p-adic invariants of geometric fibers of the universal Barsotti–Tate group on $\mathcal{M}$. Familiar examples of p-adic invariants include the p-rank and the Newton polygon. In 1999 F. Oort had the insight that if one uses "the mother of all p-adic invariants", namely the isomorphism class of the geometric fibers of the universal Barsotti–Tate group, then instead of a stratification of $\mathcal{M}$ by a finite number of subvarieties, one gets a decomposition of $\mathcal{M}$ into an infinite number of smooth locally closed subvarieties. In [30] these subvarieties are called *central leaves*, which we simplify to *leaves* here. In general the leaves in a given modular variety have moduli: They are parametrized by a scheme of finite type over k. Oort's Hecke orbit conjecture asserts that the leaves are determined by the Hecke symmetries: Every prime-to-p Hecke orbit is dense in the leaf containing it.

In this article we explain techniques motivated by the Hecke orbit problem: global ℓ-adic monodromy (Proposition 3.3), canonical coordinates on leaves (§4), hypersymmetric points (§5), local stabilizer principle (Proposition 6.1) and local rigidity (Theorem 6.2). The first is group-theoretic, while the rest four constitutes an effective "linearization method" of the Hecke orbit problem, which is illustrated in 6.3. The techniques developed so far are strong enough to affirm the Hecke orbit conjecture for Siegel modular varieties and Hilbert modular varieties. Advances in p-adic monodromy may lead to further progress; see §7.

In many ways a leaf in a modular variety $\mathcal{M}$ over a field of characteristic p resembles a Shimura variety in characteristic zero:

- The action of the Hecke symmetries on a leaf is topologically transitive according to the Hecke orbit Conjecture 3.2.
- By Proposition 3.3, the ℓ-adic monodromy of a leaf of positive dimension in $\mathcal{M}$ is $G(\mathbb{Q}_\ell)$, where G is a semisimple group attached to $\mathcal{M}$.
- A leaf is homogeneous in the sense that the local structure of a leaf are the same throughout the leaf, in view of the theory of canonical coordinates in §4.
- It seems plausible that a suitable parabolic subgroup P of an inner twist G' of G attached to a leaf $\mathcal{C}$ is closely related to the p-adic monodromy of $\mathcal{C}$; see §7.2. Hints of such a connection already appeared in [21].
- The theory of canonical coordinates suggests that one considers a leaf $\mathcal{C}$ as above to be "uniformized" by G'/P in a weak sense.

The above analogy depicts a scene in which a Shimura variety spawns an infinitude of morphed characteristic p replicas while reducing itself modulo p, an image that

resonates with the mantra of *Indra's Pearls*[1]. We hope that the readers find this analogy somewhat sound, or perhaps even pleasing.

2. Hecke symmetry on modular varieties

A modular varieties of PEL type classifies abelian varieties with a prescribed type of polarization, endomorphisms and level structure. To a given PEL type is associated a tower of modular varieties. A locally compact group, consisting of prime-to-p finite adelic points of a reductive algebraic group over $\mathbb{Q}$, operates on this tower; this action induces Hecke correspondences on a fixed modular variety in the tower.

2.1. PEL data. Let B be a finite dimensional semisimple algebra over $\mathbb{Q}$, let $\mathcal{O}_B$ be a maximal order of B maximal at p, and let $*$ be a positive involution on B preserving $\mathcal{O}_B$. Let V be a B-module of finite dimension over $\mathbb{Q}$, let $\langle\cdot,\cdot\rangle$ be a $\mathbb{Q}$-valued nondegenerate alternating form on V compatible with $(B,*)$, and let $h\colon \mathbb{C}\to \mathrm{End}_{B\otimes_{\mathbb{Q}}\mathbb{R}}(V\otimes_{\mathbb{Q}}\mathbb{R})$ be a $*$-homomorphism such that

$$(v,w)\ \mapsto\ \langle v, h(\sqrt{-1})w\rangle$$

defines a positive definite real-valued symmetric form on $V\otimes_{\mathbb{Q}}\mathbb{R}$. The 6-tuple $(B,*,\mathcal{O}_B,V,\langle\cdot,\cdot\rangle,h)$ is called a *PEL datum unramified at p* if B is unramified at p and there exists a self-dual $\mathbb{Z}_p$-lattice in $V\otimes_{\mathbb{Q}}\mathbb{Q}_p$ stable under $\mathcal{O}_B$.

2.2. Modular varieties of PEL type. Suppose that we are given a PEL datum $(B,*,\mathcal{O}_B,V,\langle\cdot,\cdot\rangle,h)$ unramified at p, one associates a tower of modular varieties $(\mathcal{M}_{K^p})$ indexed by the set of all compact open subgroups K^p of $G(\mathbb{A}_f^p)$, where G is the unitary group attached to the pair $(\mathrm{End}_B(V),*)$, and $\mathbb{A}_f^p=\prod'_{\ell\neq p}\mathbb{Q}_\ell$ is the ring of prime-to-p finite adeles attached to $\mathbb{Q}$. The modular variety $\mathcal{M}_{K^p}$ classifies abelian varieties A with endomorphisms by $\mathcal{O}_B$ plus prime-to-p polarization and level-structure, whose H_1 is modeled on the given PEL datum; see [20, §5] for details.

2.3. Hecke symmetries. The group $G(\mathbb{A}_f^p)$ operates on the whole projective system $(\mathcal{M}_{K^p})$. If a level subgroup K_0^p is fixed, then on the corresponding modular variety $\mathcal{M}_{K_0^p}$ the remnant from the action of $G(\mathbb{A}_f^p)$ takes the form of a family of algebraic finite étale algebraic correspondences on $\mathcal{M}_{K_0^p}$; they are known as *Hecke correspondences*.

For a given point $x\in\mathcal{M}_{K_0^p}(k)$, denote by $\mathcal{H}^p\cdot x$ the subset of $\mathcal{M}_{K_0^p}(k)$ consisting of all elements which belongs to the image of x under some prime-to-p Hecke correspondence on $\mathcal{M}_{K_0^p}$. The countable set $\mathcal{H}^p\cdot x$ is called the prime-to-p Hecke orbit of x; it is equal to the image of $G(\mathbb{A}_f^p)\cdot\tilde{x}$ in $\mathcal{M}_{K_0^p}(k)$, where $\tilde{x}\in\varprojlim_{K^p}\mathcal{M}_{K^p}(k)$ is a pre-image of x.

[1] Cf. the preface of [25].

2.4. Examples

2.4.1. Siegel modular varieties. Let $g, n \in \mathbb{N}$, $(n, p) = 1$, and $n \geq 3$. Denote by $\mathcal{A}_{g,n}$ the modular variety over k which classifies all g-dimensional principally polarized abelian varieties (A, λ) over k with a symplectic level-n structure η. Two k-points $[(A_1, \lambda_1, \eta_1)]$, $[(A_2, \lambda_2, \eta_2)]$ in $\mathcal{A}_{g,n}$ are in the same prime-to-p Hecke orbit if and only if there exists a prime-to-p quasi-isogeny β (="$\beta_2 \circ \beta_1^{-1}$")

$$\beta\colon A_1 \xleftarrow{\beta_1} A_3 \xrightarrow{\beta_2} A_2$$

defined by prime-to-p isogenies β_1 and β_2 such that β respects the principal polarizations λ_1 and λ_2 in the sense that $\beta_1^*(\lambda_1) = \beta_2^*(\lambda_2)$. The semisimple algebra B in the PEL datum is equal to $\mathbb{Q}$. The reductive group G attached to the PEL datum is the symplectic group Sp_{2g}. The modular variety attached to the principal congruence subgroup of level-n in $\mathrm{Sp}_{2g}(\mathbb{A}_f^p)$ is $\mathcal{A}_{g,n}$.

2.4.2. Hilbert modular varieties. Let $E = F_1 \times \cdots \times F_r$, where $F_1, \ldots, F_r$ are totally real number fields. Consider the PEL datum where $B = E$, $* = \mathrm{Id}_E$, and V is a free E-module of rank two. Then the reductive group attached to the PEL datum is the kernel of the composition

$$\prod_{E/\mathbb{Q}} \mathrm{GL}_2 \xrightarrow{\det} \prod_{E/\mathbb{Q}} \mathbb{G}_m \xrightarrow{\mathrm{Nm}_{E/\mathbb{Q}}} \mathbb{G}_m,$$

where $\prod_{E/\mathbb{Q}}$ denotes Weil's restriction of scalars functor from E to $\mathbb{Q}$. A typical member of the associated tower of modular varieties is $\mathcal{M}_{E,n}$, with $n \geq 3$ and $(n, p) = 1$, which classifies $[E : \mathbb{Q}]$-dimensional abelian varieties A over k, together with a ring homomorphism $\iota\colon \mathcal{O}_E = \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_r} \to \mathrm{End}(A)$, an $\mathcal{O}_E$-linear level-n structure η, and an $\mathcal{O}_E$-linear polarization of A.

There are different versions of polarizations; the one in [14] is as follows. It is a positivity-preserving $\mathcal{O}_E$-linear homomorphism $\lambda\colon \mathcal{L} \to \mathrm{Hom}^{\mathrm{sym}}_{\mathcal{O}_E}(A, A^t)$ from an projective rank-one $\mathcal{O}_E$-module $\mathcal{L}$ with a notion of positivity, which induces an $\mathcal{O}_E$-linear isomorphism $\lambda\colon A \otimes_{\mathcal{O}_E} \mathcal{L} \xrightarrow{\sim} A^t$, where A^t is the dual abelian variety of A. The modular variety $\mathcal{M}_{E,n}$ is not smooth over k if any one of the totally real fields F_i is ramified above p; if so then $\mathcal{M}_{E,n}$ has moderate singularities – it is a local complete intersection.

The prime-to-p Hecke orbit of a point $[(A, \iota_A, \lambda_A, \eta_A)] \in \mathcal{M}_{E,n}(k)$ consists of all points $[(B, \iota_B, \lambda_B, \eta_B)] \in \mathcal{M}_{E,n}(k)$ such that there exists a prime-to-p $\mathcal{O}_E$-linear quasi-isogeny from A to B which preserves the polarizations.

2.4.3. Picard modular varieties. Let L be an imaginary quadratic field contained in $\mathbb{C}$ such that p is unramified in L. Let a, b be positive integers, and let $g = a + b$. For the PEL datum, we take $B = L$ with the involution induced by the

complex conjugation, a g-dimensional vector space V over L, and an L-linear complex structure $h\colon \mathbb{C} \to \mathrm{End}_{L\otimes_{\mathbb{Q}}\mathbb{R}}(V\otimes_{\mathbb{Q}}\mathbb{R})$ on $V\otimes_{\mathbb{Q}}\mathbb{R}$ satisfying the following condition: For every element $u \in L \subset \mathbb{C}$, the trace of the action of u on V_1 is equal to $au + b\bar{u}$, where $V_1 = \{v \in V\otimes_{\mathbb{Q}}\mathbb{C} : h(z)(v) = z\cdot v \text{ for all } z \in \mathbb{C}\}$. We also fix a ring homomorphism $\varepsilon\colon \mathcal{O}_L \to k$, i.e. a $\mathcal{O}_L$-algebra structure on k.

Let $n \geq 3$ be a positive integer, $(n, p) = 1$. The Picard modular variety $\mathcal{M}_{L,a,b,n}$ over k classifies $\mathcal{O}_L$-linear$g$-dimensional abelian varieties (A, ι) of signature (a, b), together with a principal polarization $\lambda\colon A \to A^t$ such that $\lambda\circ\iota(\bar{u}) = \iota(u)^t\circ\lambda$ for all $u \in \mathcal{O}_L$, and a symplectic $\mathcal{O}_L$-linear level-n structure. The signature condition above is that

$$\det_{\mathcal{O}_L\otimes_{\mathbb{Z}}k}(T\cdot\mathrm{Id} - \iota(u)|\mathrm{Lie}(A,\iota)) = (T-\varepsilon(u))^a\cdot(T-\varepsilon(u))^b \in k[T] \text{ for all } u \in \mathcal{O}_L.$$

As before the prime-to-p Hecke orbit of a point $[(A, \iota_A, \lambda_A, \eta_A)] \in \mathcal{M}_{L,a,b,n}(k)$ consists of all points $[(B, \iota_B, \lambda_B, \eta_B)] \in \mathcal{M}_{L,a,b,n}(k)$ such that there exists a prime-to-p $\mathcal{O}_L$-linear quasi-isogeny from A to B which preserves the polarizations.

3. Leaves and the Hecke orbit conjecture

3.1. Leaves in modular varieties in characteristic p

Definition 3.1. Let $\mathcal{M} = \mathcal{M}_{K_0^p}$ be a modular variety of PEL type over k as in 2.2. Let x_0 be a point in $\mathcal{M}(k)$. The *leaf* $\mathcal{C}_{\mathcal{M}}(x_0)$ in $\mathcal{M}$ passing through x_0 is the reduced locally closed subvariety of $\mathcal{M}$ over k such that $\mathcal{C}_{\mathcal{M}}(x_0)(k)$ consists of all points $x = [(A, \lambda, \iota, \eta)] \in \mathcal{C}_{\mathcal{M}}(x_0)(k)$ such that the Barsotti–Tate group (or p-divisible group) $(A[p^\infty], \lambda[p^\infty], \iota[p^\infty])$ with prescribed polarization and endomorphisms attached to x is isomorphic to that attached to x_0.

Remark. (i) The notion of leaves was introduced in [30]; it was studied later by Vasiu in [38].

(ii) In addition to being locally closed, every leaf in $\mathcal{M}$ is a smooth subvariety of $\mathcal{M}$ and is closed under all prime-to-p Hecke correspondences.

3.2. The Hecke orbit conjectures. The Hecke orbit conjecture HO, due to Oort, asserts that the decomposition of a modular variety $\mathcal{M}$ of PEL type into leaves is determined by the Hecke symmetries on $\mathcal{M}$. It is equivalent to the conjunction of the continuous version $\mathrm{HO}_{\mathrm{ct}}$ and the discrete version $\mathrm{HO}_{\mathrm{dc}}$ below.

Conjecture 3.2 (HO). Every prime-to-p Hecke orbit in a modular variety of PEL type $\mathcal{M}$ over k is dense in the leaf in $\mathcal{M}$ containing it.

Conjecture ($\mathrm{HO}_{\mathrm{ct}}$). The closure of any prime-to-p Hecke orbit in the leaf $\mathcal{C}$ containing it is an open-and-closed subset of $\mathcal{C}$, i.e. it is a union of irreducible components of the smooth variety $\mathcal{C}$.

Conjecture ($\mathrm{HO}_{\mathrm{dc}}$). Every prime-to-$p$ Hecke orbit in a leaf $\mathcal{C}$ meets every irreducible component of $\mathcal{C}$.

3.3. Global ℓ-adic monodromy. The discrete Hecke orbit conjecture is essentially an irreducibility statement, in view of the following result on global monodromy.

Proposition 3.3. *Let $\mathcal{M}$ be a modular variety of PEL type attached to a reductive group G over $\mathbb{Q}$ as in* 2.2. *Let $G^{\mathrm{sc}}_{\mathrm{der}}$ be the simply connected cover of the derived group of G. Let $x_0 \in \mathcal{M}(k)$ be a point of $\mathcal{M}$ such that the prime-to-p Hecke orbit of x_0 with respect to every simple factor of $G^{\mathrm{sc}}_{\mathrm{der}}$ is infinite. Let $Z(x_0)$ be the Zariski closure of the prime-to-p Hecke orbit of x_0 for the group $G^{\mathrm{sc}}_{\mathrm{der}}$ in the leaf $\mathcal{C}(x_0)$ in $\mathcal{M}$ containing x_0. Then $Z(x_0)$ is irreducible, and the Zariski closure of the ℓ-adic monodromy group of $Z(x_0)$ is $G_{\mathrm{der}}(\mathbb{Q}_\ell)$ for every prime number $\ell \neq p$.*

Remark. (i) A stronger version of 3.3 for Siegel modular varieties is proved in [5]. The argument in [5] works for all modular varieties of PEL type.

(ii) The irreducibility statement in Proposition 3.3 is a useful tool for proving irreducibility of a given subvariety Z of modular varieties of PEL type which are stable under the prime-to-p Hecke correspondences: It reduces the task to proving Hecke transitivity on $\pi_0(Z)$.

3.4. Some known cases of the Hecke orbit conjecture

Theorem 3.4. *The Hecke orbit conjecture* HO *holds for Siegel modular varieties.*

Theorem 3.5. *The Hecke orbit conjecture* HO *holds for Hilbert modular varieties attached to a finite product $F_1 \times \cdots \times F_r$ of totally real fields. Here the prime p may be ramified in any or all of the totally real fields $F_1, \ldots, F_r$.*

Remark. (i) Theorem 3.4 is joint work with F. Oort. Details of the proof of Theorem 3.4 will appear in a monograph with F. Oort. The proof of the continuous version $\mathrm{HO}_{\mathrm{ct}}$ in the Siegel case uses Theorem 3.5.

(ii) The proof of Theorem 3.5 is the result of joint work with C.-F. Yu; the proof of the discrete version, i.e. the irreducibility of non-supersingular leaves in Hilbert modular varieties, is the work of C.-F. Yu.

(iii) Among the methods used in the proof of Theorem 3.4, the action of the local stabilizer subgroup and the trick of using Hilbert modular subvarieties first appeared in [2], where the case of Theorem 3.4 for ordinary principally polarized abelian varieties was proved.

(iv) A detailed sketch of the proof of Theorem 3.4 can be found in [4].

4. Canonical coordinates on leaves

4.1. Classical Serre–Tate coordinates. Recall that an abelian variety A over k is *ordinary* if the Barsotti–Tate group $A[p^\infty]$ is the extension of an étale Barsotti–Tate

group by a toric Barsotti–Tate group over k. It has been more than forty years when Serre and Tate discovered that the local deformation space of an ordinary abelian variety A over k has a natural structure as a formal torus over $W(k)$ of relative dimension $\dim(A)^2$. For Siegel modular varieties, their result says that if $x = [(A, \lambda)] \in \mathcal{A}_{g,n}(k)$ is a closed point of $\mathcal{A}_{g,n}$ such that A is an ordinary abelian variety, then the formal completion $\mathcal{A}_{g,n}^{/x}$ of $\mathcal{A}_{g,n} \to \mathrm{Spec}(\mathbb{Z})$ has a natural structure as a formal torus over $W(k)$ of relative dimension $\frac{g(g+1)}{2}$, where $g = \dim(A)$. Notice that the ordinary locus of $\mathcal{A}_{g,n}$ over k is equal to the dense open stratum in the Newton polygon stratification of $\mathcal{A}_{g,n}$.

The above approach generalizes to modular variety of PEL type, to the effect that the mixed-characteristic local deformation space for a point in the dense open Newton polygon stratum of a modular variety $\mathcal{M}$ of PEL type can be built up from Barsotti–Tate groups over $W(k)$ by a system of fibrations; see [23].

There is a long-standing question as to whether one can find a reasonable theory of canonical coordinates for points outside the generic Newton polygon stratum of a modular variety of PEL type. It turns out that the answer is *yes* if one restricts to a leaf in a modular variety.

4.2. The slope filtration. The starting point is the observation that there exists a natural *slope filtration* on the restriction of the universal Barsotti–Tate group to a leaf; moreover the slope filtration gives the local moduli of a leaf.

Proposition 4.1. *Let $\mathcal{M}$ be a modular variety of PEL type over k attached to a PEL datum $(B, *, \mathcal{O}_B, V, \langle\cdot,\cdot\rangle, h)$ as in* 2.2. *Let $\mathcal{C}$ be a leaf in $\mathcal{M}$. Let $\tilde{X} \to \mathcal{C}$ be the restriction to $\mathcal{C}$ of the Barsotti–Tate group attached to the universal abelian variety.*

(i) *There exist rational numbers $\mu_1, \dots \mu_m$ with $1 \geq \mu_1 > \cdots > \mu_m \geq 0$ and Barsotti–Tate groups*

$$0 = \tilde{X}_0 \subset \tilde{X}_1 \subset \tilde{X}_2 \subset \cdots \subset \tilde{X}_m = \tilde{X}$$

over $\mathcal{C}$ such that $\tilde{Y}_i := \tilde{X}_i/\tilde{X}_{i-1}$ is a non-trivial isoclinic Barsotti–Tate group over $\mathcal{C}$ of slope μ_i for each $i = 1, \dots, m$.

(ii) *The filtration $0 = \tilde{X}_0 \subset \tilde{X}_1 \subset \tilde{X}_2 \subset \cdots \subset \tilde{X}_m = \tilde{X}$ is uniquely determined by $\tilde{X} \to \mathcal{C}$. Each subgroup $\tilde{X}_i \subseteq \tilde{X}$ is stable under the natural action of $\mathcal{O}_B \otimes_{\mathbb{Z}} \mathbb{Z}_p$.*

(ii) *For each $i = 1, \dots, r$ the Barsotti–Tate group $\tilde{Y}_i \to \mathcal{C}$ is geometrically constant, hence $\tilde{Y}_i$ is isomorphic to the twist of a constant Barsotti–Tate group by a smooth étale $\mathbb{Z}_p$-sheaf over $\mathcal{C}$.*

Remark. (i) That $\tilde{Y}_i$ is *isoclinic* of slope μ_i means that the kernel of the N-th iterate of the relative Frobenius for $\tilde{Y}_i$ is comparable to the kernel of $[p^{N\mu_i}]_{\tilde{X}_i/\tilde{X}_{i-1}}$ for all sufficiently large multiples of the denominator of μ_i.

(ii) The proof of Proposition 4.1 depends on [42] and [34].

4.3. The cascade structure. Combining Proposition 4.1 with the Serre–Tate theorem, one sees that the local moduli of a leaf $\mathcal{C}$ comes from the deformation of the slope filtration.

Let $x = [(A, \iota, \lambda, \eta)] \in \mathcal{M}(k)$ be a closed point of the modular variety $\mathcal{M}$, and let $\mathrm{Fil}_{A[p^\infty]} = (0 = X_0 \subset X_1 \subset \cdots \subset X_m = A[p^\infty])$ be the slope filtration of $A[p^\infty]$. Let $Y_i = X_i/X_{i-1}$ for $i = 1, \ldots, m$. For each pair (i, j) with $1 \le i \le j \le m$, let $\mathfrak{Def}([i, j], \iota) = \mathfrak{Def}([i, j], \iota[p^\infty])$ be the deformation functor over k of the filtered Barsotti–Tate group

$$0 \subset X_i/X_{i-1} \subset X_{i+1}/X_{i-1} \subset \cdots \subset X_j/X_{i-1}$$

with action by $\mathcal{O}_B \otimes_{\mathbb{Z}} \mathbb{Z}_p$. Each $\mathfrak{Def}([i, j], \iota)$ is a smooth formal scheme over k, and $\mathfrak{Def}([i, i], \iota) = \mathrm{Spec}(k)$ for each i. For $1 \le i < j \le m$, let $\mathfrak{DE}(i, j; \iota) = \mathfrak{DE}(i, j; \iota[p^\infty])$ be the deformation functor of the filtered Barsotti–Tate group $0 \subset Y_i \subset Y_i \times_{\mathrm{Spec}(k)} Y_j$ with action by $\mathcal{O}_B \otimes_{\mathbb{Z}} \mathbb{Z}_p$; it is a smooth formal scheme over k. Each $\mathfrak{DE}(i, j; \iota)$ has a natural structure as a smooth commutative formal group over k; the group structure comes from via Baer sum. Notice that $\mathfrak{Def}([i, i+1], \iota)$ is a torsor over $\mathfrak{DE}(i, i+1; \iota])$ for $i = 1, \ldots, m-1$.

We have a family of forgetful morphisms

$$\pi_{[i+1,j],[i,j]} : \mathfrak{Def}([i, j], \iota]) \to \mathfrak{Def}([i+1, j], \iota), \quad 1 \le i < j \le m,$$

and

$$\pi_{[i,j-1],[i,j]} : \mathfrak{Def}([i, j], \iota) \to \mathfrak{Def}([i, j-1], \iota), \quad 1 \le i < j \le m$$

such that

$$\pi_{[i+1,j-1],[i+1,j]} \circ \pi_{[i+1,j],[i,j]} = \pi_{[i+1,j-1],[i,j-1]} \circ \pi_{[i,j-1],[i,j]} \quad \text{if } i \le j-2.$$

Each morphism $\pi_{[i+1,j],[i,j]}$ is smooth, same for each $\pi_{[i,j-1],[i,j]}$.

For each pair (i, j) with $1 \le i < j \le m$, define a commutative smooth formal group

$$\pi'_{[i+1,j],[i,j]} : \mathfrak{DE}(i, [i+1, j]; \iota) \to \mathfrak{Def}([i+1, j], \iota)$$

as follows. For each Artinian local ring R over k and for each R-valued point $f : \mathrm{Spec}(R) \to \mathfrak{Def}([i+1, j], \iota)$ corresponding to a deformation

$$0 \subset \tilde{X}_{[i+1,i+1]} \subset \cdots \subset \tilde{X}_{[i+1,j]}$$

of the filtration $(0 \subset X_{i+1}/X_i \subset \cdots \subset X_j/X_i)$ over R, define the set of R-valued points of $\mathfrak{DE}(i, [i+1, j], \iota)$ over f to be the set of all isomorphism classes of extensions of $\tilde{X}_{[i+1,j]}$ by $Y_i \times_{\mathrm{Spec}(k)} \mathfrak{Def}([i+1, j], \iota)$. It is easy to see that $\pi_{[i+1,j],[i,j]}$ has a natural structure as a torsor for $\pi'_{[i+1,j],[i,j]}$. Similarly one can define a commutative formal group

$$\pi'_{[i,j-1],[i,j]} : \mathfrak{DE}([i, j-1], j; \iota) \to \mathfrak{Def}([i, j-1]$$

for $1 \le i < j \le m$ so that $\pi_{[i,j-1],[i,j]}$ has a natural structure as a torsor for $\pi'_{[i,j-1],[i,j]}$.

Consider the natural map

$$\pi_{[i,j]}\colon \mathfrak{Def}([i,j],\iota) \to \mathfrak{Def}([i+1,j],\iota) \times_{\mathfrak{Def}([i+1,j-1],\iota)} \mathfrak{Def}([i,j-1],\iota)$$

defined by the maps $\pi_{[i+1,j],[i,j]}$ and $\pi_{[i,j-1],[i,j]}$. It turns out that in a suitable sense the map $\pi_{[i,j]}$ has a natural structure as a *torsor for a biextension* of

$$\big(\mathfrak{DE}([i+1,j-1],j;\iota),\ \mathfrak{DE}(i,[i+1,j-1];\iota)\big)$$

by (the base extension to $\mathfrak{Def}([i+1,j-1],\iota)$ of) the commutative smooth formal group $\mathfrak{DE}(i,j;\iota)$ if $i \le j-2$. Notice that for the two factors of the target of the map $\pi_{[i,j]}$, the first factor $\mathfrak{Def}([i+1,j],\iota) \to \mathfrak{Def}([i+1,j-1],\iota)$ is a torsor for the group $\mathfrak{DE}([i+1,j-1],j;\iota) \to \mathfrak{Def}([i+1,j-1],\iota)$, while the second factor $\mathfrak{Def}([i,j-1],\iota) \to \mathfrak{Def}([i+1,j-1],\iota)$ is a torsor for the group $\mathfrak{DE}(i,[i+1,j-1];\iota) \to \mathfrak{Def}([i+1,j-1],\iota)$.

The formal structure of a family such as

$$\begin{aligned}\mathfrak{MDE} = \big(&\mathfrak{Def}([i,j],\iota]),\ \mathfrak{Def}([i,j],\iota), \pi_{[i+1,j],[i,j]},\\ &\pi_{[i,j-1],[i,j]}, \pi'_{[i+1,j],[i,j]}, \pi'_{[i,j-1],[i,j]}, \pi_{[i,j]}\big)\end{aligned}$$

will be called a *cascade*, following the terminology in [23], although the situation here is somewhat more complicated than [23].

When x is a point of the generic Newton polygon stratum of $\mathcal{M}$, the maximal subcascade of $\mathfrak{MDE}$ fixed by the involution induced by the polarization λ coincides with the formal completion $\mathcal{M}^{/x}$ of $\mathcal{M}$ at x. So $\mathcal{M}^{/x}$ is built up from suitable subgroups of the commutative formal groups $\mathfrak{DE}(i,j;\iota[p^\infty])$ over k through a family of fibrations; see [23].

4.4. Maximal p-divisible subcascade. Suppose that x lies outside the generic Newton polygon stratum. Then when one deforms the slope filtration, the resulting Barsotti–Tate group may fail to remain geometrically constant. It turns out that the maximal reduced closed formal subscheme of $\mathfrak{Def}([0,m];\iota)$ is in some sense the *maximal p-divisible subcascade* of the cascade $\mathfrak{MDE}$ of formal groups attached to x; the latter is built up from the maximal p-divisible formal subgroups $\mathfrak{DE}(i,j;\iota)_{\mathrm{pdiv}}$ of $\mathfrak{DE}(i,j;\iota)$, with (i,j) running through all pairs with $1 \le i < j \le m$.

The polarization λ of the abelian variety A induces an involution on the formal scheme $\mathfrak{Def}([0,m];\iota])$. and also an involution of the maximal p-divisible subcascade $\mathfrak{MDE}_{\mathrm{pdiv}}$ of the cascade of formal groups $\mathfrak{MDE}$ attached to $x \in \mathcal{M}(k)$. The maximal closed formal subscheme of the formal scheme underlying $\mathfrak{MDE}_{\mathrm{pdiv}}$ is equal to the formal completion $\mathcal{C}^{/x}$ of the leaf $\mathcal{C}$ in $\mathcal{M}$ containing x. In particular $\mathcal{C}^{/x}$ is built up from p-divisible formal groups over k through a family of fibrations.

4.5. The two slope case. Let X, Y be isoclinic Barsotti–Tate groups over k with slopes $\mu_X < \mu_Y$. Let h_X and h_Y be the height of X and Y respectively. Let $\mathfrak{DE}(X, Y)$ be the deformation functor over k of the filtration $0 = Y \subset X \times_{\mathrm{Spec}(k)} Y$; it is a commutative smooth formal group over k. Let $\mathfrak{DE}(X, Y)_{\mathrm{pdiv}}$ be the maximal p-divisible formal subgroup of the commutative smooth formal group $\mathfrak{DE}(X, Y)$ over k.

Let $\mathrm{M}(X)$ and $\mathrm{M}(Y)$ be the Cartier module of X and Y respectively. We refer to [43] for the Cartier theory. On $H_{\mathbb{Q}} := \mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{\mathbb{Z}} \mathbb{Q}$ we have a σ-linear operator F and a σ^{-1}-linear operator V on $H_{\mathbb{Q}}$ given by

$$(V \cdot h)(u) = V(h(V^{-1}u)), \quad (F \cdot h)(u) = F(h(V(u)) \quad \text{for all } h \in H_{\mathbb{Q}}, u \in \mathrm{M}(X).$$

Clearly $\mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y))$ is stable under the action of F.

Theorem 4.2. *Notation as above.*

(i) *The p-divisible formal group $\mathfrak{DE}(X, Y)_{\mathrm{pdiv}}$ is isoclinic of slope $\mu_Y - \mu_X$; its height is equal to $h_X \cdot h_Y$.*

(ii) *The Cartier module of $\mathfrak{DE}(X, Y)_{\mathrm{pdiv}}$ is naturally isomorphic to the maximal $W(k)$-submodule of $\mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y))$ which is stable under the actions of F and V.*

(iii) *Suppose that $Y = X^t$ is the Serre dual of X. Then we have a natural involution $*$ on $\mathfrak{DE}(X, X^t)_{\mathrm{pdiv}}$, and the Cartier module of the maximal formal subgroup of $\mathfrak{DE}(X, X^t)_{\mathrm{pdiv}}$ fixed under $*$ is the maximal $W(k)$-submodule of $\mathrm{Hom}_{W(k)}(\mathrm{S}^2(\mathrm{M}(X)), W(k))$ which is stable under the actions of F and V.*

Remark. (i) See [8] for a proof of Theorem 4.2.

(ii) The set of all p-typical curves in the reduced Cartier ring functor, with three compatible actions by the reduced Cartier ring over k, plays a major role in the proof.

(iii) The case when we have a maximal order $\mathcal{O}_B \otimes_{\mathbb{Z}} \mathbb{Z}_p$ in an unramified semisimple algebra $B \otimes_{\mathbb{Q}} \mathbb{Q}_p$ over $\mathbb{Q}_p$ operating on X and Y is easily deducible from Theorem 4.2.

5. Hypersymmetric points

Over a field of characteristic zero one has the notion of *special points* in modular varieties of PEL type, corresponding to abelian varieties of CM-type (or, with sufficiently many complex multiplications). On the other hand, it is well-known that every abelian variety over $\overline{\mathbb{F}}_p$ has sufficiently many complex multiplications, so one can say that every $\overline{\mathbb{F}}_p$-point of a modular variety $\mathcal{M}$ of PEL type is "special". But there are points in $\mathcal{M}$ that are more distinguished than others – they correspond to abelian varieties whose $\mathcal{O}_B$-endomorphism ring is "as big as allowed by the slope constraint".

Definition 5.1. (i) Let B be a simple algebra over $\mathbb{Q}$, and let $\mathcal{O}_B$ be an order of B. Let A be an abelian variety over k, and let $\iota\colon \mathcal{O}_B \to \mathrm{End}(A)$ be a ring homomorphism. We say that (A, ι) is a *hypersymmetric $\mathcal{O}_B$-linear abelian variety* if the canonical map $\mathrm{End}_{\mathcal{O}_B}(A) \otimes_{\mathbb{Z}} \mathbb{Z}_p \to \mathrm{End}_{\mathcal{O}_B}(A[p^\infty])$ is an isomorphism.

(ii) Let $\mathcal{M}$ be a modular variety of PEL type as in 2.2. A point $x \in \mathcal{M}(k)$ is *hypersymmetric* if the underlying $\mathcal{O}_B$-linear abelian variety (A_x, ι_x) is hypersymmetric.

Remark 5.2. (i) When $B = \mathbb{Q}$, it is easy to see that an abelian variety A over k is hypersymmetric if and only if it is isogenous to a finite product of abelian varieties $B_1 \times \cdots \times B_r$ defined over a finite field $\mathbb{F}_q$ such that the action of the Frobenius element $\mathrm{Fr}_{B_i, \mathbb{F}_q}$ on the first ℓ-adic cohomology group of B_i has at most two eigenvalues for each $i = 1, \ldots, r$, and B_i and B_j share no common slope if $i \neq j$. See [9, §2, §3].

(ii) One can use the method in [9, §5] to show that there exist hypersymmetric points on any leaf of a modular variety of PEL type over k. However one has difficulty showing the existence of hypersymmetric points on every irreducible component of a given leaf of a modular variety, without knowing or assuming the irreducibility of the leaf.

(iii) If the semisimple $\mathbb{Q}$-rank of the reductive group G attached to the PEL datum for the modular variety $\mathcal{M}$ is equal to one, then every $\overline{\mathbb{F}}_p$-point of $\mathcal{M}$ is hypersymmetric. For instance, every $\overline{\mathbb{F}}_p$-point of the modular curve is hypersymmetric. One consequence of this phenomenon is that one cannot simply substitute "special points" by "hypersymmetric points" and expect to get a reasonable formulation of the André-Oort conjecture in characteristic p; see [9, §7].

6. Action of stabilizer subgroups and rigidity

6.1. Stabilizer subgroups. Let $\mathcal{M}$ be a modular variety over k of PEL type as in 2.2. Attached to a point $x \in \mathcal{M}(k)$ corresponding to a quadruple $(A_x, \iota_x, \lambda_x, \eta_x)$ are two compact p adic groups:

- Let $\mathrm{G_x}(\mathbb{Z}_p) = \mathrm{Aut}_{\mathcal{O}_B}(A_x[p^\infty], \lambda_x[p^\infty])$. We call $G_x(\mathbb{Z}_p)$ the *local p-adic automorphism group* at x, and

- Let $\mathrm{H_x}$ be the unitary group attached to the semisimple algebra with involution $(\mathrm{End}_{\mathcal{O}_B}(A_x) \otimes_{\mathbb{Z}} \mathbb{Q}, *_x)$, where $*_x$ is the Rosati involution attached to λ_x. Let $\mathrm{H}_x(\mathbb{Z}_p)$ be the group of $\mathbb{Z}_p$-points of $\mathrm{H_x}$ with respect to the integral structure given by $\mathrm{End}(A_x)$. We call $\mathrm{H}_x(\mathbb{Z}_p)$ the *local stabilizer subgroup* at x. We have a natural embedding $\mathrm{H}_x(\mathbb{Z}_p) \hookrightarrow G_x(\mathbb{Z}_p)$.

6.2. Action on deformation space. We have a natural action of $G_x(\mathbb{Z}_p)$ on $\mathcal{M}^{/x}$, the formal completion of $\mathcal{M}^{/x}$. This action comes from the combination of

(a) a classical theorem of Serre and Tate, which states that the deformation functor for the abelian variety A_x is canonically isomorphic to the deformation functor attached to the Barsotti–Tate group $A_x[p^\infty]$, and

(b) the action of $G_x(\mathbb{Z}_p)$ on the deformation functor for the $\mathcal{O}_B$-linear Barsotti–Tate group $(A_x[p^\infty], \iota_x[p^\infty])$ by "change of marking", or "transport of structure".

The local stabilizer subgroup $\mathrm{H}_x(\mathbb{Z}_p)$ can be regarded as the p-adic completion of the stabilizer subgroup at x in the set of all prime-to-p Hecke correspondences. Consequently the *local stabilizer principle* holds:

Proposition 6.1 (Local stabilizer principle). *If Z is a closed subvariety of $\mathcal{M}$ stable under all prime-to-p Hecke correspondence and $x \in Z(k)$ is a closed point of Z, then the formal completion $Z^{/x} \subset \mathcal{M}^{/x}$ of Z at x is stable under $\mathrm{H}_x(\mathbb{Z}_p)$ for the action described in* 6.2.

The local stabilizer principle can be effectively deployed for studying Hecke-invariant subvarieties when combined with the rigidity result below.

Theorem 6.2 (Local rigidity). *Let X be a p-divisible formal group over k. Let H be a connected reductive linear algebraic subgroup over $\mathbb{Q}_p$ and let $\rho\colon H(\mathbb{Q}_p) \to (\mathrm{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)^\times$ be a rational linear representation of $H(\mathbb{Q}_p)$ such that the composition of ρ with the left regular representation of $\mathrm{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ does not contain the trivial representation of $H(\mathbb{Q}_p)$ as a subquotient. Let Z be an irreducible closed formal subscheme of X. Assume that Z is stable under the natural action of an open subgroup U of $H(\mathbb{Q}_p)$ on X. Then Z is a p-divisible formal subgroup of X.*

The proof of 6.2 is elementary; see [6]. An instructive special case of 6.2 asserts that any irreducible closed formal subvariety of a formal torus over k which is stable under multiplication by $1 + p^n$ for some $n \geq 1$ is a formal subtorus.

6.3. Linearization of the Hecke orbit problem. The combination of the local stabilizer principle and local rigidity leads to an effective linearization of the Hecke orbit problem: Consider the case when $\mathcal{M}$ is a Siegel modular variety $\mathcal{A}_{g,n}$ and $x \in \mathcal{A}_{g,n}(\overline{\mathbb{F}}_p)$ corresponds to a g-dimensional principally polarized abelian variety A_x defined over $\overline{\mathbb{F}}_p$ with two slopes $\lambda < 1-\lambda$. Then the formal completion $\mathcal{C}(x)^{/x}$ at x of the leaf $\mathcal{C}(x)$ containing x has a natural structure as an isoclinic p-divisible formal group of hight $\frac{g(g+1)}{2}$ and slope $1-2\lambda$. The local stabilizer principle and Theorem 6.2 imply that the formal completion at x of the Zariski closure of the prime-to-p Hecke orbit of x is a p-divisible formal subgroup of $C(x)^{/x}$; moreover this p-divisible formal subgroup is stable under the natural action of the local stabilizer subgroup $\mathrm{H}_x(\mathbb{Z}_p)$.

Continuing the situation above, and assume that A_x is hypersymmetric in the sense of 5.1. Then the Zariski closure of the prime-to-p Hecke orbit of x coincides with the irreducible component of $\mathcal{C}(x)$ in an open neighborhood of x, because the action on the local stabilizer subgroup $\mathrm{H}_x(\mathbb{Z}_p)$ on the Cartier module of the p-divisible formal group $C(x)^{/x}$ underlies an absolutely irreducible representation of $\mathrm{H}_x(\mathbb{Q}_p)$.

6.4. Hypersymmetric points and the Hecke orbit conjecture. Let $x \in \mathcal{M}(\overline{\mathbb{F}}_p)$ be an $\overline{\mathbb{F}}_p$ point of a modular variety $\mathcal{M}$ of PEL type, and let $Z(x)$ be the Zariski closure in the leaf $\mathcal{C}(x)$ of the prime-to-p Hecke orbit of x. The argument in 6.3 shows that the continuous Hecke orbit conjecture $\mathrm{HO}_{\mathrm{ct}}$ for $Z(x)$ would follow if one can show that there exists a hypersymmetric point y in $Z(x)$.

The Hecke orbit conjecture HO for Hilbert modular varieties comes into the proof of the continuous Hecke orbit conjecture $\mathrm{HO}_{\mathrm{ct}}$ for Siegel modular varieties at this juncture. After a possibly inseparable isogeny correspondence, one can assume that the given point $x \in \mathcal{A}_{g,n}$ lies in a Hilbert modular subvariety in $\mathcal{A}_{g,n}$. After another application of the local stabilizer principle, one is reduced to the case when the abelian variety A_x has only two slopes. With the help of Theorem 3.5, one sees that $Z(x)$ contains the leaf through x in a Hilbert modular subvariety containing x, hence $Z(x)$ contains a hypersymmetric point in $\mathcal{A}_{g,n}$. See [4] for a more detailed outline of the argument.

Remark 6.3. (i) The proof of the Hecke orbit conjecture for Siegel modular varieties outlined above relies on a special property of Siegel modular varieties: For every point $x \in \mathcal{A}_{g,n}(\overline{\mathbb{F}}_p)$, there exists a Hilbert modular variety $\mathcal{M}_{E,n}$, a finite-to-one correspondence $f\colon \mathcal{M}_{E,n} \to \mathcal{A}_{g,n}$ equivariant with respect to the prime-to-p Hecke correspondences, and a point $y \in \mathcal{M}_{E,n}(\overline{\mathbb{F}}_p)$ above x. See [4, §9], labeled as the "Hilbert trick". The point is that, every $\overline{\mathbb{F}}_p$-point of $\mathcal{A}_{g,n}$ lies in a subvariety which is essentially the reduction of a "small" Shimura subvariety of positive dimension, namely a Hilbert modular variety attached to a product E of totally real fields such that $\dim_{\mathbb{Q}}(E) = g$. Here "small" means that every factor of the reductive group attached to the Shimura subvariety has semisimple $\mathbb{Q}$-rank one.

(ii) The property that every rational point over a finite field lies in the image of a small Shimura variety of positive dimension holds for modular varieties of PEL type C, but fails for modular varieties of type A and D. Consequently, the Hecke orbit conjecture for modular varieties of PEL type C is within reach by available methods, while new ideas are needed for PEL types A and D, or the reduction of general Shimura varieties.

7. Open questions and outlook

We discuss two approaches toward a proof of the Hecke orbit conjecture for Siegel modular varieties without resorting to the Hilbert trick.

7.1. Tate-linear subvarieties in leaves. For simplicity, we consider the special case of a leaf $\mathcal{C}$ in a Siegel modular variety $\mathcal{A}_{g,n}$, such that every point of $\mathcal{C}$ corresponds to a g-dimensional principally polarized abelian variety with two slopes $\lambda < 1 - \lambda$. This assumption implies that the formal completion $\mathcal{C}^{/x}$ of $\mathcal{C}$ has a natural structure

as an isoclinic p-divisible formal group with slope $1-2\lambda$ and height $\frac{g(g+1)}{2}$, for any closed point $x \in \mathcal{C}$.

An irreducible closed subvariety $Z \subset \mathcal{C}$ is said to be *Tate linear* at a closed point $x \in Z$ if the formal completion $Z^{/x}$ is a p-divisible formal subgroup of $\mathcal{C}^{/x}$. It can be shown that if Z is Tate-linear at one closed point of $\mathcal{C}$, then it is Tate-linear at every closed point of the smooth locus of Z

Remark. (i) The proof that the property of being Tate-linear propagates from one point of Z to every point of Z depends on a global version of canonical coordinates. The case when $\mathcal{C}$ is the ordinary locus of $\mathcal{A}_{g,n}$ has been documented in [7], where several issues related to the notion of Tate-linear subvarieties are addressed.

(ii) The notion of Tate-linear subvarieties is inspired by the Hecke orbit problem: Suppose that $\mathcal{M}$ is a modular variety of PEL type contained in a Siegel modular variety $\mathcal{A}_{g,n}$, and $x \in \mathcal{M}(\overline{\mathbb{F}}_p) \cap \mathcal{C}(\overline{\mathbb{F}}_p)$ is a point of $\mathcal{M}$ such that the abelian variety A_x attached to x has two slopes $\lambda < 1-\lambda$. Then the Zariski closure of the Hecke orbit $\mathcal{H}^p \cdot x$ in the leaf $\mathcal{C}_{\mathcal{M}}(x)$ is a Tate linear subvariety of $\mathcal{C}$.

Question 7.1. The most intriguing question about the notion of Tate-linear subvarieties is whether every Tate-linear subvariety of a leaf $\mathcal{C}$ in $\mathcal{A}_{g,n}$ is (an irreducible component of) the intersection of $\mathcal{C}$ with the reduction of a Shimura subvariety of $\mathcal{A}_{g,n}$ in characteristic 0.

It seems plausible that the answer is a *qualified yes*. This naive expectation will be termed the *global rigidity conjecture*.

Remark. (i) If the global rigidity conjecture is true, then the notion of Tate-linear subvarieties provides a geometric characterization for subvarieties of $\mathcal{C}$ which are equal to (an irreducible component of) the intersection of $\mathcal{C}$ with the reduction of a Shimura subvariety of $\mathcal{A}_{g,n}$.

(ii) The global rigidity conjecture should be considered as being stronger than the continuous Hecke orbit conjecture $\mathrm{HO}_{\mathrm{ct}}$: Continuing the set-up as in §7.1. Let Z be the Zariski closure in the leaf $\mathcal{C}_{\mathcal{M}}(x)$ in $\mathcal{M}$ containing x of the prime-to-p Hecke orbit in $\mathcal{M}$. Then Z is a Tate-linear subvariety, by Theorem 6.2. Moreover if the global rigidity conjecture is true, then one can deduce without difficulty that Z is a union of irreducible components of $\mathcal{C}(x)$.

7.2. *p*-adic monodromy. As we saw in 6.3, the combination of the local stabilizer principle, canonical coordinates and the local rigidity theory achieves a certain level of localization for the Hecke orbit problem. This linearization allows one to approach the Hecke orbit problem through the p-adic monodromy: Let Z be the Zariski closure of a given prime-to-p Hecke orbit $\mathcal{H}(x)$ in the leaf $\mathcal{C}$ containing $\mathcal{H}(x)$. Consider the restriction to Z of the universal Barsotti–Tate group $A[p^\infty] \to \mathcal{M}$ over the modular variety $\mathcal{M}$. Over the leaf $\mathcal{C}$, the Barsotti–Tate group $A[p^\infty] \to \mathcal{C}$ admits a *slope filtration*; the p-adic monodromy attached to the associated graded of the slope filtration of $A[p^\infty] \to \mathcal{Z}$ will be called the *naive p-adic monodromy* of $A[p^\infty] \to Z$.

Conjecture 7.2. The naive p-adic monodromy of the family $A[p^\infty] \to Z$ is "as large as possible", in the sense that the image of the naive p-adic monodromy representation is an open subgroup of the group of $\mathbb{Q}_p$-points of a suitable Levi subgroup L of an inner twist G' of G attached to Z, where G is the reductive group attached to the PEL data for $\mathcal{M}$.

Remark. (i) Conjecture 7.2 for the Zariski closure Z of an Hecke orbit implies the continuous Hecke orbit conjecture $\mathrm{HO}_{\mathrm{cont}}$.

(ii) Conjecture 7.2 is a p-adic analogue of Proposition 3.3.

(iii) As a weak converse to Conjecture 7.2, the method of the proof of Proposition 7.4 below should enable one to show that the Hecke orbit conjecture HO implies Conjecture7.2, using a hypersymmetric point as the base point.

Remark 7.3. Given an abelian scheme $A \to S$, where S a scheme over $\mathbb{F}_p$, we would like to show that the naive p-adic monodromy for $A \to S$ is "as large as possible", subject to obvious constraints, such as cycles on the family $A \to S$. As an intermediate step toward this goal, one would like to show that, when S is the spectrum of a Noetherian local integral domain and the Newton polygon of the closed fiber of $A \to S$ is different from the Newton polygon of the generic fiber, the naive p-adic monodromy for the generic fiber of $A \to S$ is *large* in a suitable sense.

When $\dim(A/S) = 1$, the above *wish* is a classical theorem of Igusa. The argument of Igusa was generalized in [3] to the case of a one-dimensional p-divisible formal group with ordinary generic fiber. The same argument applies to the case of a p-divisible formal group with ordinary generic fiber such that the dimension and the codimension are coprime; details will appear in an article with D. U. Lee.

Proposition 7.4. *Let $\mathcal{A}_{g,n}^{\mathrm{or}}$ be the ordinary locus of a Siegel modular variety $\mathcal{A}_{g,n}$ over k, where $g \geq 1$, $n \geq 3$, $(n, p) = 1$, and the base field $k \supseteq \mathbb{F}_p$ is algebraically closed. Let $A \to \mathcal{A}_{g,n}^{\mathrm{or}}$ be the universal abelian scheme over $\mathcal{A}_{g,n}^{\mathrm{or}}$. Let $A[p^\infty]_{\mathrm{et}} \to \mathcal{A}_{g,n}^{\mathrm{or}}$ be the maximal étale quotient of $A[p^\infty] \to \mathcal{A}_{g,n}^{\mathrm{or}}$; it is an étale Barsotti–Tate group of height g. Let E_0 be an ordinary elliptic curve defined over $\overline{\mathbb{F}}_p$, and let $x_0 = (A_0, \lambda_0)$, where A_0 is the product of g copies of E_0, and λ_0 is the product principal polarization on A_0. Let $\mathrm{T}_p = \mathrm{T}_p(A_0[p^\infty])_{\mathrm{et}})$ be the p-adic Tate module of the étale p-divisible group $A_0[p^\infty]_{\mathrm{et}}$; it is naturally isomorphic to the direct sum of g copies of $\mathrm{T}_p(E_0[p^\infty]_{\mathrm{et}}) \cong \mathbb{Z}_p$, so $\mathrm{GL}(\mathrm{T}_p)$ is naturally isomorphic to $\mathrm{GL}_g(\mathbb{Z}_p)$. Let $\rho\colon \pi_1(\mathcal{A}_{g,n}^{\mathrm{or}}, x_0) \to \mathrm{GL}(\mathrm{T}_p)$ be the naive p-adic monodromy representation of $A[p^\infty] \to \mathcal{A}_{g,n}^{\mathrm{or}}$. Then the image of ρ is equal to $\mathrm{GL}(\mathrm{T}_p) \cong \mathrm{GL}_g(\mathbb{Z}_p)$.*

Proof. Let $\mathcal{B}$ be the product of g copies of $\mathcal{A}_{1,n}$, diagonally embedded in $\mathcal{A}_{g,n}$. Let E_0 be an ordinary elliptic curve defined over a $\overline{\mathbb{F}}_p$, and let $x_0 = (A_0, \lambda_0)$, where A_0 is the product of g copies of E_0, and λ_0 is the product principal polarization on A_0. Let $\mathcal{O} = \mathrm{End}(E_0)$. Then $\mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Z}_p \cong \mathbb{Z}_p \times \mathbb{Z}_p$, corresponding to the natural splitting of $E_0[p^\infty]$ into the product of its toric part $E_0[p^\infty]_{\mathrm{tor}}$ and its étale part $E_0[p^\infty]_{\mathrm{et}}$. So we have an isomorphism $\mathrm{End}(A_0) \cong \mathrm{M}_g(\mathcal{O})$, and a splitting $\mathrm{End}(A_0) \otimes_{\mathbb{Z}} \mathbb{Z}_p \cong$

$\mathrm{M}_g(\mathcal{O}) \times \mathrm{M}_g(\mathcal{O})$ corresponding to the splitting of $A_0[p^\infty]$ into the product its toric and étale parts. Denote by $\mathrm{pr}\colon (\mathrm{End}(A_0) \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times \to \mathrm{GL}(\mathrm{T}_p) \cong \mathrm{GL}_g(\mathbb{Z}_p)$ the projection corresponding to the action of $\mathrm{End}(A_0) \otimes_{\mathbb{Z}} \mathbb{Z}_p$ on the étale factor $A_0[p^\infty]_{\mathrm{et}}$ of $A_0[p^\infty]$. The Rosati involution $*$ on $\mathrm{End}(A_0)$ interchanges the two factors of $\mathrm{End}(A_0) \otimes_{\mathbb{Z}} \mathbb{Z}_p$. It follows that $\mathrm{U}(\mathcal{O}_{(p)} \otimes_{\mathbb{Z}} \mathbb{Z}_p, *)$ is isomorphic to $\mathrm{GL}(\mathrm{T}_p)$ under the projection map pr, therefore the image of $\mathrm{U}(\mathcal{O}_{(p)}, *)$ in $\mathrm{GL}(\mathrm{T}_p)$ is dense in $\mathrm{GL}(\mathrm{T}_p)$. Here $\mathcal{O}_{(p)} = \mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Z}_{(p)}$, and $\mathbb{Z}_{(p)} = \mathbb{Q} \cap \mathbb{Z}_p$ is the localization of $\mathbb{Z}$ at the prime ideal $(p) = p\mathbb{Z}$.

By a classical theorem of Igusa, the p-adic monodromy group of the restriction to $\mathcal{B}$, i.e. $\rho(\mathrm{Im}(\pi_1(\mathcal{B}, x_0) \to \pi_1(\mathcal{A}_{g,n}, x_0)))$, is naturally identified with the product of g copies of $\mathbb{Z}_p^\times$ diagonally embedded in $\mathrm{GL}(\mathrm{T}_p) \cong \mathrm{GL}_g(\mathbb{Z}_p)$. Denote by D this subgroup of $\mathrm{GL}(\mathrm{T}_p)$.

Let $R_{(p)} = \mathrm{End}(A_0) \otimes_{\mathbb{Z}} \mathbb{Z}_{(p)} \cong \mathrm{M}_g(\mathcal{O}) \otimes_{\mathbb{Z}} \mathbb{Z}_{(p)}$. Every element $u \in R_{(p)}$ such that $u^*u = uu^* = 1$ gives rise to a prime-to-p isogeny from A_0 to itself respecting the polarization λ_0. Such an element $u \in R_{(p)}$ gives rise to

- a prime-to-p Hecke correspondence h on $\mathcal{A}_{g,n}$ having x_0 as a fixed point, and
- an irreducible component $\mathcal{B}'$ of the image of $\mathcal{B}$ under h such that $\mathcal{B}' \ni x_0$.

By the functoriality of the fundamental group, the image of the fundamental group $\pi_1(\mathcal{B}', x_0)$ of $\mathcal{B}'$ in $\pi_1(\mathcal{A}^{\mathrm{or}}_{g,n}, x_0)$ is mapped under the p-adic monodromy representation ρ to the conjugation of D by the element $\mathrm{pr}(h) \in \mathrm{GL}(\mathrm{T}_p)$. In particular, $\rho(\pi_1(\mathcal{A}^{\mathrm{or}}_{g,n}, x_0))$ is a closed subgroup of $\mathrm{GL}(\mathrm{T}_p)$ which contains all conjugates of D by elements in the image of $\mathrm{pr}\colon \mathrm{U}(E_{(p)}, *) \to \mathrm{GL}(\mathrm{T}_p)$.

Recall that the image of $\mathrm{U}(E_{(p)}, *)$ in $\mathrm{GL}(\mathrm{T}_p)$ is a dense subgroup. So the monodromy group $\rho(\pi_1(\mathcal{A}_{g,n}, x_0))$ is a closed normal subgroup of $\mathrm{GL}(\mathrm{T}_p) \cong \mathrm{GL}_g(\mathbb{Z}_p)$ which contains the subgroup D of all diagonal elements. An easy exercise in group theory shows that the only such closed normal subgroup is $\mathrm{GL}_g(\mathbb{Z}_p)$ itself. □

Remark. (i) There are at least two published proofs of Proposition 7.4 in the literature, in [15] and [16, chap. V §7] respectively.

(ii) In the proof of 7.4, one can use as the base point any element $[(A_1, \lambda)]$ of $\mathcal{A}^{\mathrm{or}}_{g,n}$ such that A_1 is separably isogenous to a product of g copies of an ordinary elliptic curve E_1 over $\overline{\mathbb{F}}_p$.

(iii) As already mentioned before, the argument of Proposition 7.4 applies to leaves in modular varieties of PEL type. Since we used a hypersymmetric point as the base point, a priori this argument applies only to those irreducible components of a given leaf which contain hypersymmetric points.

References

[1] Andreatta, F., Goren, E. Z., Hilbert modular varieties of low dimension. In *Geometric Aspects of Dwork's Theory* (ed. by A. Adolphson, F. Baldassarri, P. Berthelot, N. Katz, and F. Loeser), Volume I, Walter de Gruyter, Berlin 2004, 113–175.

[2] Chai, C.-L., Every ordinary symplectic isogeny class in positive characteristic is dense in the moduli. *Invent. Math.* **121** (1995), 439–479.

[3] Chai, C.-L., Local monodromy for deformations of one-dimensional formal groups. *J. Reine Angew. Math.* **524** (2000), 227–238.

[4] Chai, C.-L., Hecke orbits on Siegel modular varieties. In *Geometric Methods in Algebra and Number Theory*, Progr. Math. 235, Birkhäuser, Boston, MA, 2004, 71–107.

[5] Chai, C.-L., Monodromy of Hecke-invariant subvarieties. *Quaterly J. Pure Applied Math.* **1** (Borel Special Issues, part I) (2005), 291–303.

[6] Chai, C.-L., A rigidity result for p-divisible formal groups. Preprint, 2003. Available from http://www.math.upenn.edu/~chai/.

[7] Chai, C.-L., Families of ordinary abelian varieties: canonical coordinates, p-adic monodromy, Tate-linear subvarieties and Hecke orbits. Preprint, 2003.

[8] Chai, C.-L., Canonical coordinates on leaves of p-divisible groups: The two-slope case. Preprint, 2004.

[9] Chai, C.-L., Oort, F., Hypersymmetric abelian varieties. *Quaterly J. Pure Applied Math.* **2** (1) (2006) (Coates Special Issue), 1–27.

[10] Chai, C.-L., Oort, F., *Hecke Orbits*. Monograph in preparation.

[11] Chai, C.-L., Yu, C.-F., Fine structures and Hecke orbits on Hilbert modular varieties. Preliminary, 2004; revision in progress.

[12] de Jong, A. J., Homomorphisms of Barsotti-Tate groups and crystals in positive characteristic. *Invent. Math.* **134** (1998), 301–333.

[13] de Jong, A. J., Oort F., Purity of the stratification by Newton polygons, *J. Amer. Math. Soc.* **13** (2000), 209–241.

[14] P. Deligne and G. Pappas, Singularités des espaces de modules de Hilbert, en les caractéristiques divisant le discriminant. *Composito Math.* **90** (1994), 59–79.

[15] Ekedahl, T., The action of monodromy on torsion points of jacobian. In *Arithmetic Algebraic Geometry*, Progr. Math. 89, Birkhäuser, Boston, MA, 1991, 41–49.

[16] Faltings, G. and Chai, C.-L., *Degeneration of Abelian Varieties*. Ergeb. Math. Grenzgeb. (3) 22, Springer-Verlag, Berlin 1990.

[17] Goren, E. Z., Oort, F., Stratification of Hilbert modular varieties. *J. Algebraic Geom.* **9** (2000), 111–154.

[18] Grothendieck, I., *Groupes de Monodromie en Géométrie Algébrique* (SGA 7) I. Lecture Notes in Math. 288, Springer-Verlag, Berlin1972.

[19] Katz, N. M., Slope filtration of F-crystals. *Astérisque* **63** (1979), 113–164.

[20] Kottwitz, R. E., Points on some Shimura varieties over finite fields. *J. Amer. Math. Soc.* **2** (1992), 373–444.

[21] Kottwitz, R. E., Isocrystals with additional structures. *Compositio Math.* **56** (1985), 201–220.

[22] Manin, Y. I., The theory of commutative formal groups over fields of finite characteristic. *Uspehi Mat. Nauk* **1**8 (1963), 3–90; English transl. *Russian Math. Surveys* **18** (1963), 1–80.

[23] Moonen, B., Serre-Tate theory for moduli spaces of PEL type. *Ann. Sci. École Norm. Sup.* **37** (2004), 223–269.

[24] Mumford, D., Bi-extensions of formal groups. In *Algebraic Geometry* (Bombay, 1968), Oxford University Press, London 1969, 307–322.

[25] Mumford, D., Series, C., Wright, D., *Indra's Pearls. The Vision of Felix Klein*. Cambridge University Press, Cambridge 2002.

[26] Oda, T., Oort, F., Supersingular abelian varieties. In *Proceedings of the International Symposium on Algebraic Geometry* (Kyoto, 1977), Kinokuniya Book Store, Tokyo 1978, 595–621.

[27] Oort, F., Some questions in algebraic geometry. Preliminary version, 1995; available from http://www.math.uu.nl/people/oort/.

[28] Oort, F., Newton polygons and formal groups: conjectures by Manin and Grothendieck. *Ann. of Math.* **152** (2000), 183–206.

[29] Oort, F., A stratification of a moduli space of abelian varieties. In *Moduli of Abelian Varieties*, Progr. Math. 195, Birkhäuser, Basel 2001, 345–416.

[30] Oort, F., Foliations in moduli spaces of abelian varieties. *J. Amer. Math. Soc.* **17** (2004), 267–296.

[31] Oort, F., Minimal p-divisible groups. *Ann. of Math.* **161** (2005), 1021–1036

[32] Oort, F., Monodromy, Hecke orbits and Newton polygon strata. Seminar at Max Planck Institut, Bonn, Feb. 14, 2004; available from http://www.math.uu.nl/people/oort/.

[33] Oort, F., Hecke orbits in moduli spaces. Notes for the 2005 AMS Summer Institute on Algebraic Geometry, 22 pp.

[34] Oort, F., Zink, T., Families of p-divisible groups with constant Newton polygon. *Doc. Math.*, to appear.

[35] Ribet, K., p-adic interpolation via Hilbert modular forms. *Algebraic Geometry* (Arcata 1974), Proc. Symp. Pure Math. 29, Amer. Math. Soc., Providence, R.I., 1975, 581–592.

[36] Tate, J., Endomorphisms of abelian varieties over finite fields. *Invent. Math.* **2** (1966), 134–144.

[37] Tate, J., Classes d'isogenie de variétés abéliennes sur un corps fini (d'après T. Honda). In *Séminaire Bourbaki 1968/69*, Exposé 352, Lecture Notes in Math. 179, Springer-Verlag, Berlin 1971, 95–110.

[38] Vasiu, A., Crystalline boundedness principle. Preprint, 2005; available from http://math.arizona.edu/~vasiu/.

[39] Yu, C.-F., On the supersingular locus in the Hilbert-Blumenthal 4-folds. *J. Algebraic Geom.* **12** (2003), 653–698.

[40] Yu, C.-F., On reduction of Hilbert-Blumenthal varieties. *Ann. Inst. Fourier* **53** (2003), 2105–2154.

[41] Yu, C.-F., Discrete Hecke orbit problem for Hilbert-Blumenthal varieties. Preprint, 2004.

[42] Zink, T., On the slope filtration. *Duke Math. J.* **109** (2001), 79–95.

[43] Zink, T., Cartiertheorie kommutativer formaler Gruppen. Teubner-Texte Math. 68, B. G. Teubner Verlagsgesellschaft, Leipzig 1984.

Department of Mathematics, University of Pennsylvania, 209 S. 33rd Street, Philadelphia, PA 19104-6395, U.S.A.

E-mail: chai@math.upenn.edu

Heegner points, Stark–Heegner points, and values of L-series

Henri Darmon*

Abstract. Elliptic curves over $\mathbb{Q}$ are equipped with a systematic collection of *Heegner points* arising from the theory of complex multiplication and defined over abelian extensions of imaginary quadratic fields. These points are the key to the most decisive progress in the last decades on the Birch and Swinnerton-Dyer conjecture: an essentially complete proof for elliptic curves over $\mathbb{Q}$ of analytic rank ≤ 1, arising from the work of Gross–Zagier and Kolyvagin. In [Da2], it is suggested that Heegner points admit a host of conjectural generalisations, referred to as *Stark–Heegner points* because they occupy relative to their classical counterparts a position somewhat analogous to Stark units relative to elliptic or circular units. A better understanding of Stark–Heegner points would lead to progress on two related arithmetic questions: the explicit construction of global points on elliptic curves (a key issue arising in the Birch and Swinnerton-Dyer conjecture) and the analytic construction of class fields sought for in Kronecker's Jugendtraum and Hilbert's twelfth problem. The goal of this article is to survey Heegner points, Stark–Heegner points, their arithmetic applications and their relations (both proved, and conjectured) with special values of L-series attached to modular forms.

Mathematics Subject Classification (2000). Primary 11G05; Secondary 11G15.

Keywords. Elliptic curves, modular forms, L-series, Heegner points, Stark–Heegner points.

1. Introduction

Elliptic curves are distinguished among projective algebraic curves by the fact that they alone are endowed with the structure of a (commutative) algebraic group. The *affine* curves with this property are the additive group $\mathbb{G}_a$ and the multiplicative group $\mathbb{G}_m$. The integral points on $\mathbb{G}_a$ (taken, say, over an algebraic number field F) is a finitely generated $\mathbb{Z}$-module. The same is true for the integral points on G_m: these are the units of F, whose structure is well understood thanks to Dirichlet's unit theorem. The close parallel between units and rational points on elliptic curves is frequently illuminating. In both cases, it is the natural group law on the underlying curve which lends the associated Diophantine theory its structure and richness.

An elliptic curve E over F can be described concretely as a Weierstrass equation

*The author is grateful to NSERC, CICMA, McGill University and the Centre de Recherches Mathématiques in Montreal for their support during the writing of this paper.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

in projective space

$$y^2z = x^3 + axz^2 + bz^3, \quad a, b \in F, \text{ where } \Delta := 4a^3 - 27b^2 \neq 0.$$

The group $E(F)$ of F-rational (or equivalently: integral) solutions to this equation is in bijection with the F-rational solutions of the corresponding affine equation

$$y^2 = x^3 + ax + b,$$

together with an extra "point at infinity" corresponding to $(x, y, z) = (0, 1, 0)$.

The most basic result on the structure of $E(F)$ is the *Mordell–Weil Theorem* which asserts that $E(F)$ is a finitely generated abelian group, so that there is an isomorphism of abstract groups

$$E(F) \simeq T \oplus \mathbb{Z}^r,$$

where T is the finite torsion subgroup of $E(F)$. The integer $r \geq 0$ is called the *rank* of E over F. Many questions about T are well-understood, for example:

1. There is an efficient algorithm for computing T, given E and F.

2. A deep result of Mazur [Ma] describes the possible structure of T when $F = \mathbb{Q}$ and E is allowed to vary over all elliptic curves. The size of T is bounded uniformly, by 14. Mazur's result has been generalised by Kamienny and Merel [Mer], yielding a uniform bound on the size of T when F is fixed – a bound which depends only on the degree of F over $\mathbb{Q}$.

In contrast, much about the rank remains mysterious. For example, can r become arbitrarily large, when F is fixed but E is allowed to vary? The answer is believed to be yes, but no proof is known for $F = \mathbb{Q}$ or for any other number field F.

An even more fundamental problem resides in the absence of effectivity in the proof of the Mordell–Weil theorem. Specifically, the answer to the following question is not known.

Question 1.1. Is there an algorithm which, given E, calculates the rank r of $E(F)$, and a system $P_1, \ldots, P_r$ of generators for this group modulo torsion?

A candidate for such an algorithm is Fermat's method of infinite descent, but this method is not guaranteed to terminate in a finite amount of time – it would, if the so-called *Shafarevich–Tate group* $Ш(E/\mathbb{Q})$ of E is finite, as is predicted to be the case.

Question 1.1 is also connected with the *Birch and Swinnerton-Dyer conjecture*. This conjecture relates Diophantine invariants attached to E, such as r, to the Hasse–Weil L-series $L(E, s)$ of E, a function of the complex variable s which is defined in terms of an Euler product taken over the non-archimedean places v of F. To describe this Euler product precisely, let $\mathbb{F}_v = \mathcal{O}_F/v$ denote the residue field of F at v, and write $|v| := \#\mathbb{F}_v$ for the norm of v. The elliptic curve E is said to have *good reduction*

at v if it can be described by an equation which continues to describe a smooth curve over $\mathbb{F}_v$ after reducing its coefficients modulo v. Set $\delta_v = 1$ if E has good reduction at v, and $\delta_v = 0$ otherwise. Finally, define integers a_v indexed by the places v of good reduction for E by setting

$$a_v := |v| + 1 - \#E(\mathbb{F}_v).$$

This definition is extended to the finite set of places of bad reduction for E, according to a recipe in which $a_v \in \{0, 1, -1\}$, the precise value depending on the type of bad reduction of E in an explicit way.

The L-series of E is given in terms of these invariants by

$$L(E, s) = \prod_v (1 - a_v|v|^{-s} + \delta_v|v|^{1-2s})^{-1} = \sum_{\mathfrak{n}} a_E(\mathfrak{n})|\mathfrak{n}|^{-s},$$

where the product is taken over all the non-archimedean places v of F, and the sum over the integral ideals $\mathfrak{n}$ of F. The Euler product converges absolutely for $\mathrm{Re}(s) > 3/2$, but $L(E, s)$ is expected to admit an analytic continuation to the entire complex plane. Some reasons for this expectation, and a statement of the Birch and Swinnerton-Dyer conjecture, are given in Section 2.6.

2. Elliptic curves over $\mathbb{Q}$

It is useful to first discuss elliptic curves over $\mathbb{Q}$, a setting in which a number of results currently admit more definitive formulations.

Given an elliptic curve $E/\mathbb{Q}$, let N denote its *conductor*. This positive integer, which measures the arithmetic complexity of E, is divisible by exactly the same primes as those dividing the minimal discriminant of E (the minimum being taken over all possible plane cubic equations describing E). Denote by a_n the coefficient of n^{-s} in the Hasse–Weil L-series of E:

$$L(E, s) = \prod_p (1 - a_p p^{-s} + \delta_p p^{1-2s})^{-1} = \sum_{n=1}^{\infty} a_n n^{-s}.$$

2.1. Modular parametrisations. Little can be asserted about the effective determination of $E(\mathbb{Q})$, or about the analytic behaviour of $L(E, s)$, without the knowledge that E is *modular*. Wiles's far-reaching program for proving the modularity of elliptic curves (and more general Galois representations) has been completely carried out in [BCDT] when $F = \mathbb{Q}$. One way of formulating the modularity of E is to state that the generating series

$$f_E(z) := \sum_{n=1}^{\infty} a_n e^{2\pi i n z} \tag{1}$$

is a *modular form of weight* 2 for the Hecke congruence group

$$\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \text{ such that } N|c \right\}.$$

This means that $f(z)$ is a holomorphic function on the Poincaré upper half-plane

$$\mathcal{H} := \{z = x + iy,\ y > 0\} \subset \mathbb{C},$$

satisfying

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 f(z) \quad \text{for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N), \tag{2}$$

together with suitable growth properties around the fixed points of parabolic elements of $\Gamma_0(N)$. These fixed points belong to $\mathbb{P}_1(\mathbb{Q})$, and it is useful to replace $\mathcal{H}$ by the completed upper half-plane $\mathcal{H}^* := \mathcal{H} \cup \mathbb{P}_1(\mathbb{Q})$. After suitably defining the topology and complex structure on the quotient $\Gamma_0(N)\backslash\mathcal{H}^*$, thus making it into a compact Riemann surface, the differential form $\omega_f := 2\pi i f(z)dz$ is required to extend to a holomorphic differential on this surface.

The quotient $\Gamma_0(N)\backslash\mathcal{H}^*$ can even be identified with the set of complex points of an algebraic curve defined over $\mathbb{Q}$, denoted by $X_0(N)$. This algebraic curve structure arises from the interpretation of $\Gamma_0(N)\backslash\mathcal{H}$ as classifying isomorphism classes of elliptic curves with a distinguished cyclic subgroup of order N, in which the orbit $\Gamma_0(N)\tau \in \Gamma_0(N)\backslash\mathcal{H}$ is identified with the pair $\left(\mathbb{C}/\langle 1, \tau\rangle, \left\langle \frac{1}{N} \right\rangle\right)$. A (highly singular, in general) equation for $X_0(N)$ as a plane curve over $\mathbb{Q}$ is given by the polynomial $G_N(x, y)$ of bidegree $\#\mathbb{P}_1(\mathbb{Z}/N\mathbb{Z})$, where

$$G_N(x, y) \in \mathbb{Q}[x, y] \quad \text{satisfies} \quad G_N(j(\tau), j(N\tau)) = 0, \tag{3}$$

and j is the classical modular function of level 1.

An equivalent formulation of the modularity property is that there exists a non-constant map of algebraic curves defined over $\mathbb{Q}$,

$$\Phi_E : X_0(N) \longrightarrow E, \tag{4}$$

referred to as the *modular parametrisation* attached to E. One of the attractive features of this modular parametrisation is that it can be computed by analytic means, without the explicit knowledge of an equation for $X_0(N)$ as an algebraic curve over $\mathbb{Q}$. (Such an equation, as in (3), tends to be complicated and difficult to work with numerically for all but very small values of N.)

To describe Φ_E analytically, i.e., as a map

$$\Phi_E^{\infty} : X_0(N)(\mathbb{C}) = \Gamma_0(N)\backslash\mathcal{H} \longrightarrow E(\mathbb{C}), \tag{5}$$

let $\Lambda_f \subset \mathbb{C}$ be the set of complex numbers of the form

$$\int_{\tau}^{\gamma\tau} \omega_f, \quad \text{for } \gamma \in \Gamma.$$

It can be shown that Λ_f is a lattice, and that the quotient $\mathbb{C}/\Lambda_f$ is isomorphic to an elliptic curve E_f which is defined over $\mathbb{Q}$ and is $\mathbb{Q}$-isogenous to E. (The curve E_f is sometimes called the *strong Weil curve* attached to f.) The modular parametrisation to E_f, denoted by Φ_f, is defined analytically by the rule

$$\Phi_f(\tau) = \int_{i\infty}^{\tau} 2\pi i f(z)dz = \sum_{n=1}^{\infty} \frac{a_n}{n} e^{2\pi i n\tau} \pmod{\Lambda_f}, \tag{6}$$

for all $\tau \in \Gamma_0(N)\backslash\mathcal{H} \subset X_0(N)(\mathbb{C})$. The resulting value is viewed as an element of $E_f(\mathbb{C})$ via the identification $\mathbb{C}/\Lambda_f = E_f(\mathbb{C})$.

After choosing an isogeny $\alpha\colon E_f \longrightarrow E$ defined over $\mathbb{Q}$, the parametrisation Φ_E is defined by setting $\Phi_E^{\infty} = \alpha\Phi_f$. In practice it is preferable to start with $E = E_f$, at the cost of replacing E by a curve which is isogenous to it, so that α can be chosen to be the identity. The map Φ_E^{∞} is then given directly by (6).

2.2. Heegner points. Let $K \subset \mathbb{C}$ be a quadratic imaginary field, and denote by K^{ab} its maximal abelian extension, equipped with an embedding into $\mathbb{C}$ compatible with the complex embedding of K. The following theorem, a consequence of the theory of complex multiplication, is one of the important applications of the modular parametrisation Φ_E of (5):

Theorem 2.1. *If τ belongs to $K \cap \mathcal{H}$, then $\Phi_E^{\infty}(\tau)$ belongs to $E(K^{\mathrm{ab}})$.*

Theorem 2.1 also admits a more precise formulation which describes the field of definition of $\Phi_E^{\infty}(\tau)$. Let $M_0(N) \subset M_2(\mathbb{Z})$ denote the ring of 2×2 matrices with integer entries which are upper triangular modulo N. Given $\tau \in \mathcal{H}$, the *associated order* of τ is the set of matrices in $M_0(N)$ which preserve τ under Möbius transformations, together with the zero matrix, i.e.,

$$\mathcal{O}_\tau := \left\{\gamma \in M_0(N) \text{ such that } \gamma\begin{pmatrix}\tau\\1\end{pmatrix} = \lambda_\gamma\begin{pmatrix}\tau\\1\end{pmatrix}, \text{ for some } \lambda_\gamma \in \mathbb{C}\right\}.$$

The assignment $\gamma \mapsto \lambda_\gamma$ identifies $\mathcal{O}_\tau$ with a discrete subring of $\mathbb{C}$. Such rings are isomorphic either to $\mathbb{Z}$, or to an order in a quadratic imaginary field, the latter case occurring precisely when τ generates a quadratic (imaginary) extension of $\mathbb{Q}$. In that case $\mathcal{O}_\tau$ is an order in the quadratic field $K = \mathbb{Q}(\tau)$.

Orders in quadratic fields have the peculiarity that they are completely determined by their discriminants. Write D for the discriminant of the order $\mathcal{O} = \mathcal{O}_\tau$, and let $G_D := \mathrm{Pic}(\mathcal{O})$ denote the class group of this order, consisting of isomorphism classes of projective modules of rank one over $\mathcal{O}$ equipped with the group law arising from the tensor product. A standard description identifies G_D with a quotient of the idèle class group of K:

$$G_D = \mathbb{A}_K^{\times}/\Big(K^{\times}\mathbb{C}^{\times}\mathbb{A}_{\mathbb{Q}}^{\times}\prod_{\ell}\mathcal{O}_\ell^{\times}\Big). \tag{7}$$

Here $\mathbb{A}_K^\times$ denotes the group of idèles of K, the product is taken over the rational primes ℓ, and $\mathcal{O}_\ell := \mathcal{O} \otimes \mathbb{Z}_\ell$. The group G_D also admits a more classical description which is well adapted to explicit computations, as the set of equivalence classes of primitive binary quadratic forms of discriminant D equipped with the classical Gaussian composition law. (For more details on this classical point of view, see Bhargava's lecture in these proceedings.)

If D and N are relatively prime, and $\mathcal{O}_\tau = \mathcal{O}_D$, there is a primitive integral binary quadratic form $F_\tau(x, y) = A_\tau x^2 + B_\tau xy + C_\tau y^2$ satisfying

$$F_\tau(\tau, 1) = 0, \quad B_\tau^2 - 4A_\tau C_\tau = D, \quad N \text{divides } A_\tau.$$

In particular,

$$\text{all the primes } \ell | N \text{ are split in } K/\mathbb{Q}, \tag{8}$$

and therefore the equation

$$x^2 = D \pmod{N}$$

has a solution (namely, B_τ). Fix a square root δ of D modulo N, and define

$$\mathcal{H}^D := \{\tau \in \mathcal{H} \text{ such that } \mathcal{O}_\tau = \mathcal{O}_D \text{ and } B_\tau \equiv \delta \pmod{N}\}.$$

The function which to $\tau \in \Gamma_0(N)\backslash\mathcal{H}^D$ associates the $\mathrm{SL}_2(\mathbb{Z})$-equivalence class of the binary quadratic form F_τ is a bijection. (Cf., for example, Section I.1 of [GKZ].) Through this bijection, $\Gamma_0(N)\backslash\mathcal{H}^D$ inherits a natural action of G_D via the Gaussian composition law. Denote this action by $(\sigma, \tau) \mapsto \tau^\sigma$, for $\sigma \in G_D$ and $\tau \in \Gamma_0(N)\backslash\mathcal{H}^D$.

Class field theory identifies G_D with the Galois group of an abelian extension of K, as is most readily apparent, to modern eyes, from (7). This abelian extension, denoted by H_D, is called the *ring class field* attached to $\mathcal{O}$, or to the discriminant D. When D is a fundamental discriminant, H_D is Hilbert class field of K, i.e., the maximal unramified abelian extension of K. Let

$$\mathrm{rec}\colon G_D \longrightarrow \mathrm{Gal}(H_D/K) \tag{9}$$

denote the reciprocity law map of global class field theory.

A more precise form of Theorem 2.1 is given by

Theorem 2.2. *If τ belongs to $\Gamma_0(N)\backslash\mathcal{H}^D$, then $\Phi_E^\infty(\tau)$ belongs to $E(H_D)$, and*

$$\Phi_E^\infty(\tau^\sigma) = \mathrm{rec}(\sigma)^{-1}\Phi_E^\infty(\tau), \quad \textit{for all } \sigma \in G_D.$$

The fact that Φ_E^∞ intertwines the explicit action of G_D on $\Gamma_0(N)\backslash\mathcal{H}^D$ arising from Gaussian composition with the natural action of $\mathrm{Gal}(H_D/K)$ on $E(H_D)$ gives a concrete realisation of the reciprocity map (9) of class field theory. It is a special case of the *Shimura reciprocity law*.

The points $\Phi_E^\infty(\tau)$, as τ ranges over $\mathcal{H} \cap K$ are called *Heegner points* attached to K. (Sometimes, this appellation is confined to the case where the discriminant of $\mathcal{O}_\tau$ is relatively prime to N.) Theorems 2.1 and 2.2 are of interest for the following reasons, which are discussed at greater length in Sections 2.3, 2.4, and 2.5 respectively.

1. They provide a simple, computationally efficient construction of rational and algebraic points on E.

2. They are a manifestation of the fact that we dispose of an *explicit class field theory* for imaginary quadratic fields, allowing the construction of abelian extensions of such fields from values of modular functions evaluated at quadratic imaginary arguments of the upper half-plane.

3. There are deep connections between the points $\Phi_E^\infty(\tau)$, for $\tau \in \mathcal{H}^D$, and the first derivative at $s = 1$ of the Hasse–Weil L-series $L(E/K, s)$ and of related partial L-series associated to ideal classes of K. These connections lead to new insights into the behaviour of these L-series and the Birch and Swinnerton-Dyer conjecture.

2.3. The efficient calculation of global points. The fact that the theory of complex multiplication, combined with modularity, can be used to construct rational and algebraic points on E is of interest in its own right. This was noticed and exploited by Heegner, and taken up systematically by Birch in the late 60s and early 70s [BS], [Bi].

Given any (not necessarily fundamental) discriminant D for which $\mathcal{H}^D \neq \emptyset$, let $K = \mathbb{Q}(\sqrt{D})$ and set

$$P_D := \mathrm{trace}_{H_D/\mathbb{Q}}(\Phi_E^\infty(\tau)), \quad \text{for any } \tau \in \mathcal{H}^D,$$
$$P_K := \mathrm{trace}_{H/K}(\Phi_E^\infty(\tau)), \quad \text{for any } \tau \in \mathcal{H}^D,\ D = \mathrm{Disc}(K).$$

When are the points P_D and P_K of *infinite order* (in $E(\mathbb{Q})$ and $E(K)$ respectively)? This question is part of the larger problem of efficiently constructing rational or algebraic points of infinite order on elliptic curves. It is instructive to consider this problem from the point of view of its *computational complexity*.

From the outset, one is stymied by the fact that an answer to Question 1.1 is not known. Complexity issues are therefore better dealt with by focussing on the following more special problem, which depends on the curve E and a positive real parameter h. To state this problem precisely, define the *height* of a rational number $r = a/b$ (represented, of course, in lowest terms) to be

$$\mathrm{height}(r) = \log(|ab| + 1).$$

Thus, the height of r is roughly proportional to the number of digits needed to write r down. The height of an equation is the sum of the heights of its coefficients. The height of a solution $(x_1, \ldots, x_n)$ to such an equation is taken to be the sum of the height of the x_j. (In the case of an elliptic curve, one might prefer a coordinate-free definition by taking the height of E to be the height of the minimal discriminant of E.)

It is expected that, for infinitely many E, the smallest height of a point of infinite order in $E(\mathbb{Q})$ can be at least as large as an exponential function of the height of E. In this respect, the behaviour of elliptic curves is not unlike that of Pell's equation,

where a fundamental solution to $x^2 - Dy^2 = 1$ has height roughly $O(\sqrt{D})$ if $\mathbb{Q}(\sqrt{D})$ has class number one. Of greatest relevance for complexity questions are the "worst-case" elliptic curves E for which the point of infinite order $P_{\min}$ of smallest height in $E(\mathbb{Q})$ has height which is *large* relative to the height of E, i.e., for which

$$\text{height}(P_{\min}) \gg \exp(\text{height}(E)).$$

In order to focus on these curves, and avoid technical side-issues associated with elliptic curves having non-torsion points of small height, we formulate the following problem:

Problem 2.3. Given an elliptic curve E, and a real number

$$h > \exp(\text{height}(E)), \tag{10}$$

find a point P of infinite order on E with $\text{height}(P) < h$, if it exists, or assert that no such point exists, otherwise.

Denote by $P(E, h)$ the instance of this problem associated to E and the parameter h. In light of (10), this parameter can be chosen as a natural measure of the size of the problem.

Note that $P(E, h)$ continues to make sense for *any* Diophantine equation. Even in such great generality, problem $P(E, h)$ has the virtue of possessing an algorithmic solution: a brute force search over all possible points (in the projective space in which E is embedded) of height less than h, say. Such an exhaustive search requires $O(\exp(h))$ operations to solve an instance of $P(E, h)$. The exponential complexity of the brute force approach provides a crude benchmark against which to measure other approaches, and leads naturally to the following definition.

Definition 2.4. A class $\mathcal{C}$ of Diophantine equations is said to be *solvable in polynomial time* if there exists $n \in \mathbb{N}$ and an algorithm that solves $P(E, h)$, with $E \in \mathcal{C}$, in at most $O(h^n)$ operations.

The property that $\mathcal{C}$ is solvable in polynomial time can be expressed informally by stating that the time required to find a *large* solution to any $E \in \mathcal{C}$ is *not much worse* than the time is takes to write that solution down. Thus, an (infinite) class $\mathcal{C}$ of equations being solvable in polynomial time indicates that there is a method for "zeroing in" on a solution (x_0, y_0) to any equation in $\mathcal{C}$ in a way that is qualitatively more efficient than running though all candidates of smaller height.

The prototype for a class of equations that possess a polynomial time solution in the sense of Definition 2.4 is Pell's equation. A polynomial time algorithm for finding a fundamental solution to $x^2 - Dy^2 = 1$ is given by the continued fraction method that was known to the Indian mathematicians of the 10th century (although Fermat seems to be the first to have shown its effectivity.) See [Le] for a more thorough discussion of Pell's equation from the point of view of its computational complexity.

The strong analogy that exists between Pell's equation and elliptic curves suggests that the class **ELL** of all elliptic curves over $\mathbb{Q}$ might also be solvable in polynomial time. Indeed, Fermat's method of infinite descent (applied, say, to a rational 2-isogeny η, if it exists) reduces $P(E, h)$ to d_E instances of $P(C_1, h/2), \ldots, P(C_{d_E}, h/2)$ where the C_j are principal homogeneous spaces for E, and the number d_E is related to the cardinality of the Selmer group attached to η. Applying this remark iteratively suggests that the complexity for solving $P(E, h)$ might be a polynomial of degree related to d_E. The analysis required to make this discussion precise does not appear in the literature, and it would be interesting to determine whether the method of infinite descent can be used to determine to what extent **ELL** is solvable in polynomial time (assuming, eventually, the finiteness of the Shafarevich–Tate group of an elliptic curve).

It should be stressed that the method of descent is often complicated in practice because of the mounting complexity of the principal homogeneous spaces that arise in the procedure. On the other hand, the Heegner point construction, *when it produces a point of infinite order in* $E(\mathbb{Q})$, can be used to solve $P(E, h)$ by a method that is also extremely efficient in practice. See [El2] for a discussion of this application of the Heegner point construction.

For example, let

$$E : y^2 + y = x^3 - x^2 - 10x - 20$$

be the strong Weil curve of conductor 11. (This is the elliptic curve over $\mathbb{Q}$ of smallest conductor.) The following table lists a few values of the x-coordinate of P_K for some more or less randomly chosen K. It takes a desktop computer a fraction of a second to find these x-coordinates, far less than would be required to find points of comparable height on the corresponding quadratic twist of E by a naive search.

$\mathrm{Disc}(K)$	$x(P_K)$
-139	$\frac{-208838\sqrt{-139}-3182352}{1957201}$
-211	$\frac{-11055756376\sqrt{-211}-36342577392}{29444844025}$
-259	$\frac{64238721198\sqrt{-259}-2458030017103}{992886694969}$
-1003	$\frac{-2420904161556151656963 8\sqrt{-1003}-10531813107543863542748 47}{219167070502034515453609}$

2.4. Explicit Class Field Theory. The Heegner point construction is a manifestation of an explicit class field theory for imaginary quadratic fields. Normally, this is stated in terms of the elliptic modular function j. The field

$$K^? := \bigcup_{\alpha\in\mathbb{Q},\,\tau\in K\cap\mathcal{H}} K(e^{2\pi i\alpha},\, j(\tau))$$

obtained by adjoining to the imaginary quadratic field K all the roots of unity, as well as the values $j(\tau)$ for $\tau \in K \cap \mathcal{H}$, is *almost* equal to the maximal abelian extension K^{ab} of K. More precisely, $K^{\mathrm{ab}}/K^{?}$ is an extension whose Galois group, although infinite, has exponent two. (See [Se].)

Given a negative (not necessarily fundamental) discriminant D, let $\tau_1, \ldots, \tau_h$ be representatives for $\mathcal{H}^D$ (with $N = 1$) modulo the action of $\mathrm{SL}_2(\mathbb{Z})$. Then the so-called *modular polynomial*

$$Z_D(z) := \prod_{i=1}^{h} (z - j(\tau_i)) \tag{11}$$

has rational coefficients and its splitting field is the ring class field attached to the discriminant D. One might also fix an elliptic curve E and consider the function $j_E(\tau)$ of $\tau \in \Gamma_0(N)\backslash\mathcal{H}^D$ defined as the x-coordinate of the point $\Phi_E^{\infty}(\tau)$, where the x coordinate refers, say, to a minimal Weierstrass equation for E. Let Z_D^E denote the polynomial defined as in (11) with j replaced by j_E.

For example, consider the discriminants $D = -83$, -47, and -71 of class number 3, 5 and 7 respectively. The polynomials Z_D attached to the first two of these discriminants are given by:

$$x^3 + 2691907584000x^2 - 41490055168000000x + 549755813888000000000$$

$$x^5 + 2257834125x^4 - 9987963828125x^3 + 5115161850595703125x^2 - 14982472850828613281250x + 160429296006238708496093750.$$

(The degree seven polynomial Z_{-71} has been omitted to save space, its coefficients being integers of roughly 30 digits.) The following table gives the values of the polynomials $Z_D^E(z)$ for a few elliptic curves (labelled according to the widely used conventions of the tables of Cremona [Cr2]) whose conductor is a prime that splits in $\mathbb{Q}(\sqrt{D})$, for these three discriminants.

E	$Z^E_{-83}(x)$	$Z^E_{-47}(x)$
$37A$	$x^3 + 5x^2 + 10x + 4$	$x^5 - x^4 + x^3 + x^2 - 2x + 1$
$61A$	$x^3 - 2x^2 + 2x + 1$	$x^5 - x^3 + 2x^2 - 2x + 1$
$79A$		$x^5 + 4x^4 + 3x^3 - 3x^2 - x + 1$

E	$Z^E_{-71}(x)$
$37A$	$x^7 - 2x^6 + 9x^5 - 10x^4 - x^3 + 8x^2 - 5x + 1$
$43A$	$x^7 + 2x^6 + 2x^5 + x^3 + 3x^2 + x + 1$
$79A$	$x^7 + 4x^6 + 5x^5 + x^4 - 3x^3 - 2x^2 + 1$

This data illustrates the well-known fact that in computing class fields one is often better off working with modular functions other than j (such as modular units for instance). The above data suggests (at least anecdotally) that the functions j_E can be excellent choices in certain cases. For a systematic discussion of the heights of Heegner points and of the polynomials $Z_D^E(x)$ as D varies, see [RV].

2.5. Relation with L-series. The following result of Gross and Zagier [GZ] provides a connection between Heegner points and the L-series of E over K.

Theorem 2.5. *The height of P_K is equal to an explicit non-zero multiple of $L'(E/K, 1)$.*

In particular, the point P_K is of infinite order if and only if $L'(E/K, 1) \neq 0$. This result can be exploited in two ways.

Firstly, since Heegner points are so readily computable, specific instances where the point P_K is of finite order yield non-trivial examples where $L'(E/K, 1) = 0$. The *vanishing* of the leading term in an L-series is notoriously difficult to prove numerically. The Gross–Zagier theorem makes it possible to produce elliptic curves for which, provably, $L(E, 1) = L'(E, 1) = 0$. Considerations involving the sign in the functional equation for $L(E, s)$ may even force this function to vanish to odd order, and therefore to order at least 3, at $s = 1$. (The smallest elliptic curve of prime conductor with this property has conductor 5077.) The existence of elliptic curves and modular forms whose L-series has a triple zero at $s = 1$ was exploited to great effect by Goldfeld [Go] in his effective solution of the analytic class number problem of Gauss.

Secondly, and more germane to the theme of this survey, the Gross–Zagier theorem gives a criterion for the "Heegner point method" to produce a point of infinite order on $E(K)$ or on $E(\mathbb{Q})$. This provides a neat characterization of the elliptic curves for which Heegner points lead to an efficient solution of problem $P(E, h)$.

When $\mathrm{ord}_{s=1}(L(E, s)) \geq 2$, constructing the Mordell–Weil group $E(\mathbb{Q})$ is more elusive. It is an apparent paradox of the subject that we are the least well-equipped to produce global points on elliptic curves in precisely those cases when these points are expected to be more plentiful! (On the other hand, this reflects a common occurrence in mathematics, where an object that is uniquely defined is easier to produce explicitly.)

2.6. The Birch and Swinnerton-Dyer conjecture. The Birch and Swinnerton-Dyer conjecture relates the behaviour of $L(E, s)$ at $s = 1$ to arithmetic invariants of E over $\mathbb{Q}$, such as its rank. To facilitate the subsequent exposition, we state it in a form that involves an integer parameter $r \geq 0$.

Conjecture 2.6 (BSD_r). If $\mathrm{ord}_{s=1} L(E, s) = r$, then the rank of $E(\mathbb{Q})$ is equal to r, and the Shafarevich–Tate group $Ш(E/\mathbb{Q})$ of E is finite.

The Birch and Swinnerton-Dyer conjecture predicts that $E(\mathbb{Q})$ should be infinite precisely when $L(E, 1) = 0$. (The latter condition can be easily ascertained computationally in examples, because $L(E, 1)$ is known *a priori* to belong to a specific sublattice of $\mathbb{R}$.)

Remark 2.7. The Birch and Swinnerton-Dyer conjecture (suitably generalised) is consistent with the presence of a systematic supply of algebraic points defined over certain ring class fields of imaginary quadratic fields. To elucidate this remark, we begin by noting that the Birch and Swinnerton-Dyer conjecture generalises to elliptic

curves over number fields, where it predicts that the rank of $E(F)$ is equal to the order of vanishing of $L(E/F, s)$ at $s = 1$. This L-series (and its twists $L(E/F, \chi, s)$ by abelian characters of $\mathrm{Gal}(\bar{F}/F)$) admits a functional equation relating $L(E/F, \chi, s)$ to $L(E/F, \bar{\chi}, 2 - s)$. Suppose that E is defined over $\mathbb{Q}$, that $F = K$ is a quadratic extension of $\mathbb{Q}$, and that $\chi : \mathrm{Gal}(H/K) \longrightarrow \mathbb{C}^\times$ factors through the Galois group of a ring class field H of K. Then the definition of $L(E/K, \chi, s)$ as an Euler product shows that

$$L(E/K, \chi, s) = L(E/K, \bar{\chi}, s).$$

The *sign* that appears in the functional equation of the L-series $L(E/K, \chi, s)$, denoted by $\mathrm{sign}(E/K, \chi) \in \{-1, 1\}$, therefore determines the parity of its order of vanishing $\mathrm{ord}_{s=1}(L(E/K, \chi, s))$.

When (E, K) satisfies the *Heegner hypothesis* of equation (8), it can be shown that $\mathrm{sign}(E, K) = -1$ so that $L(E/K, 1) = 0$. Moreover, the same is true of $\mathrm{sign}(E/K, \chi)$ when χ is *any* ring class character of conductor prime to N_E, so that $L(E/K, \chi, 1) = 0$ for such ring class characters. In particular, if H is a ring class field of K of discriminant prime to N_E, we find

$$\mathrm{ord}_{s=1} L(E/H, s) = \mathrm{ord}_{s=1} \Big(\prod_{\chi \in \widehat{\mathrm{Gal}(H/K)}} L(E/K, \chi, s) \Big) \geq [H : K], \tag{12}$$

so that the Birch and Swinnerton-Dyer conjecture predicts the inequality:

$$\mathrm{rank}(E(H)) \overset{?}{\geq} [H : K]. \tag{13}$$

The Gross–Zagier formula (Theorem 2.5), suitably generalised to the L-series $L(E/K, \chi, s)$ with character, as in the work of Zhang discussed in Section 3.4, makes it possible to bound the rank of $E(H)$ from below by establishing the non-triviality of certain Heegner points, and yields

Corollary 2.8. *If the inequality in* (12) *is an* equality*, then the inequality* (13) *holds.*

A short time after the proof of the Gross–Zagier formula, Kolyvagin discovered a general method for using Heegner points to bound the ranks of Mordell–Weil groups *from above.*

Theorem 2.9 (Kolyvagin). *If P_K is of infinite order, then $E(K)$ has rank one and $Ш(E/K)$ is finite.*

Crucial to Kolyvagin's proof is the fact that the Heegner point P_K does not come alone, but is part of an infinite collection of algebraic points

$$\{\Phi_E^\infty(\tau)\}_{\tau \in \mathcal{H}^D}$$

as D ranges over all discriminants of orders in K. These points are defined over abelian extensions of K and obey precise compatibility relations under the norm

maps. They are used to construct a supply of cohomology classes that can be used, under the non-triviality assumption on P_K, to bound $E(K)$ and $Ш(E/K)$, showing that the former has rank one and the latter is finite. See [Ko] (or the expositions given in [Gr3] or Chapter X of [Da2]) for the details of the argument.

In relation with Corollary 2.8 we note the following consequence of Theorem 2.9 (suitably adapted to the problem of bounding Mordell–Weil groups over ring class fields in terms of Heegner points, as in [BD1] for example)

Corollary 2.10. *If the inequality in* (12) *is an* equality*, then the inequality predicted in* (13) *is an* equality.

Theorem 2.9 completes Theorem 2.5 by relating the system of Heegner points attached to E/K to the arithmetic of E over K. When combined with Theorem 2.5, it yields the following striking evidence for the Birch and Swinnerton-Dyer conjecture.

Theorem 2.11. *Conjectures BSD_0 and BSD_1 are true for all elliptic curves over $\mathbb{Q}$.*

Sketch of proof. If $\mathrm{ord}_{s=1} L(E, s) \leq 1$, one can choose an auxiliary quadratic imaginary field K in which all the primes dividing N are split, and for which

$$\mathrm{ord}_{s=1} L(E/K, s) = 1.$$

The existence of such a K is a consequence of non-vanishing results for special values and derivatives of twisted L-series. (See the book [MM], for example, for an attractive exposition of these results.) After choosing such a K, Theorem 2.5 implies that P_K is of infinite order, since $L'(E/K, 1) \neq 0$. Theorem 2.9 then implies that P_K generates a finite index subgroup of $E(K)$, and that $Ш(E/K)$ is finite. Explicit complementary information on the action of $\mathrm{Gal}(K/\mathbb{Q})$ on the point P_K implies that the rank of $E(\mathbb{Q})$ is at most one, with equality occurring precisely when $L(E, 1) = 0$. The finiteness of $Ш(E/K)$ directly implies the finiteness of $Ш(E/\mathbb{Q})$ since the restriction map $Ш(E/\mathbb{Q}) \longrightarrow Ш(E/K)$ has finite kernel. □

Theorem 2.11 is the best evidence at present for Conjecture 2.6. We remark that almost nothing is known about this conjecture when $r > 1$.

3. Elliptic curves over totally real fields

Summarising the discussion of the previous chapter, the Heegner point construction (attached to an elliptic curve over $\mathbb{Q}$, and a quadratic imaginary field K) is appealing because it provides an elegant and efficient method for calculating global points on elliptic curves as well as class fields of imaginary quadratic fields. It also leads to a proof of Conjecture BSD_r for $r = 0$ and 1.

It is therefore worthwhile to investigate whether elliptic curves defined over a number field F other than $\mathbb{Q}$ are equipped with a similar collection of algebraic

points. The modularity property so crucial in defining Heegner points does have an analogue for elliptic curves defined over F, which is most conveniently couched in the language of automorphic representations: an elliptic curve E/F should correspond to an automorphic representation π of $\mathrm{GL}_2(\mathbb{A}_F)$, the correspondence being expressed in terms of an equality of associated L-series:

$$L(E, s) = L(\pi, s).$$

(For an explanation of these concepts, see for example [Ge] or [BCdeSGKK].)

When $F = \mathbb{Q}$, the automorphic form attached to E corresponds to a differential on a modular curve, and leads to the modular parametrisation Φ_E^∞ of (4). Unfortunately, such a geometric formulation of modularity is not always available; therefore the Heegner point construction does not carry over to other number fields without further ideas.

The number fields for which Heegner points are best understood are the *totally real fields*. Let F be such a field, of degree ν, and fix an ordering $v_1, \ldots, v_\nu$ on the real embeddings of F. For $x \in F$, write $x_j := v_j(x)$ $(1 \leq j \leq \nu)$. The v_j determine an embedding of F into $\mathbb{R}^\nu$ and an embedding of $\mathrm{SL}_2(\mathcal{O}_F)$ as a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})^\nu$ with finite covolume. Given any ideal $\mathcal{N}$ of $\mathcal{O}_F$, denote by $\Gamma_0(\mathcal{N})$ the subgroup of $\mathrm{SL}_2(\mathcal{O}_F)$ consisting of matrices which are upper-triangular modulo $\mathcal{N}$.

Assume now for simplicity that F has *narrow class number one*. (The definitions to be made below need to be modified in the general case, by adopting adèlic notation which is better suited to working in greater generality but might also obscure the analogy with the classical case that we wish to draw.) A *Hilbert modular form* of parallel weight 2 and level $\mathcal{N}$ is a holomorphic function $f(z_1, \ldots, z_\nu)$ on $\mathcal{H}^\nu$ satisfying the transformation rule analogous to (2), for all matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(\mathcal{N})$:

$$f\left(\frac{a_1 z_1 + b_1}{c_1 z_1 + d_1}, \ldots, \frac{a_\nu z_\nu + b_\nu}{c_\nu z_\nu + d_\nu}\right) = (c_1 z_1 + d_1)^2 \ldots (c_\nu z_\nu + d_\nu)^2 f(z_1, \ldots, z_\nu), \quad (14)$$

together with suitable growth properties around the fixed points of parabolic elements of $\Gamma_0(\mathcal{N})$, which imply in particular that f admits a Fourier expansion "near infinity"

$$f(z_1, \ldots, z_\nu) := a(0) + \sum_{\mathfrak{n} \gg 0} a(\mathfrak{n}) e(\delta^{-1}\mathfrak{n} \cdot z),$$

in which the sum is taken over all totally positive $\mathfrak{n} \in \mathcal{O}_F$, δ is a totally positive generator of the different ideal of F, and

$$e(\mathfrak{n} \cdot z) := \exp(2\pi i (\mathfrak{n}_1 z_1 + \cdots + \mathfrak{n}_\nu z_\nu)).$$

Let $N \in F$ be a totally positive generator of the conductor ideal of E over F, and let $a_E(\mathfrak{n})$ denote the coefficients in the Hasse–Weil L-series of this elliptic curve. The following conjecture is a generalisation of the Shimura–Taniyama–Weil conjecture for totally real fields (of narrow class number one)

Conjecture 3.1. The generating series analogous to (1)

$$f_E(z_1, \ldots, z_v) := \sum_{\mathfrak{n} \gg 0} a_E(\mathfrak{n}) e(\delta^{-1}\mathfrak{n} \cdot z)$$

is a modular form of parallel weight 2 and level $\mathcal{N}$.

The methods of Wiles have successfully been extended to prove many instances of Conjecture 3.1, under a number of technical hypotheses. (See [SW] [Fu] for example.) In the sequel, it will always be assumed that any elliptic curve E/F satisfies the conclusion of Conjecture 3.1, to avoid having to worry about the precise technical conditions under which this is known unconditionally. (These conditions are fluid and ever-changing, and one might hope that they will eventually be completely dispensed with. This hope is bolstered by the wealth of new ideas – which the reader can appreciate, for instance, by consulting [BCDT], [SW], or [Ki], to cite just three in a roster that is too long and rapidly evolving to give anything like a complete list – emerging from the branch of number theory devoted to generalising and extending the scope of Wiles's methods.)

The differential form $\omega_f := f(z_1, \ldots, z_v)dz_1 \ldots dz_v$ defines a $\Gamma_0(\mathcal{N})$-invariant holomorphic differential on $\Gamma_0(\mathcal{N})\backslash\mathcal{H}^v$, but these objects do not give rise to a modular parametrisation. (Indeed, the natural generalisation of modular curves are Hilbert modular varieties, which are of dimension $[F : \mathbb{Q}]$ and probably do not admit any non-constant maps to E when $F \neq \mathbb{Q}$.) To define Heegner points on $E(F)$, it becomes crucial to consider *Shimura curve* parametrisations arising from automorphic forms on certain quaternion algebras.

3.1. Shimura curve parametrisations. Let S be a set of places of odd cardinality, containing all the archimedean places of F. Associated to S there is a *Shimura curve* denoted by X_S. This curve has a canonical model over F arising from a connection between it and the solution to a moduli problem classifying abelian varieties with "quaternionic endomorphisms". (Cf. Section 1.1 of [Zh1], for example, where it is called M_K.)

For each place $v \in S$, the curve X_S also admits an explicit v-adic analytic description. Since this description is useful for doing concrete calculations with X_S, we now describe it in some detail, following a presentation that the author learned from Gross. (Cf. [Gr4].)

If $v \in S$ is an archimedean (and hence, real) place, denote by $\mathcal{H}_v$ the Poincaré upper half-plane. If v is non-archimedean, let $\mathbb{C}_v$ denote the completion of the algebraic closure of F_v, and let $\mathcal{H}_v := \mathbb{P}_1(\mathbb{C}_v) - \mathbb{P}_1(F_v)$ denote the v-adic upper half-plane. It is equipped with a natural structure as a v-adic analytic space which plays the role of the complex structure on $\mathcal{H}$ in the non-archimedean case. (See for instance Chapter IV of [Da2] for a description of this structure.)

Let B denote the quaternion algebra over F which is ramified precisely at the places of $S - \{v\}$. (Since this set of places has even cardinality, such a quaternion

algebra exists; it is unique up to isomorphism.) Identifying v with the corresponding embedding $F \longrightarrow F_v$ of F into its completion at v, there is an F_v-algebra isomorphism

$$\iota_v \colon B \otimes_v F_v \longrightarrow M_2(F_v).$$

Let R denote a maximal $\mathcal{O}_F$-order of B if v is archimedean, and a maximal $\mathcal{O}_F[1/v]$-order of B if v is non-archimedean, and write $R_1^\times$ for the group of elements of R of reduced norm 1. Then $\Gamma_v := \iota_v(R_1^\times)$ is a discrete and finite covolume (and cocompact, if $(F, S) \neq (\mathbb{Q}, \infty)$) subgroup of $\mathrm{SL}_2(F_v)$. The quotient $\Gamma\backslash\mathcal{H}_v$ is naturally equipped with the structure of a complex curve (if v is real) or of a rigid analytic curve over $\mathbb{C}_v$ (if v is non-archimedean).

Theorem 3.2. *The quotient $\Gamma\backslash\mathcal{H}_v$ is analytically isomorphic to $X_S(\mathbb{C}_v)$.*

The complex uniformisation of $X_S(\mathbb{C})$ at the real places of F follows directly from the description of X_S in terms of the solution to a moduli problem. The non-archimedean uniformisation follows from the theory of Cerednik and Drinfeld. For more details on Drinfeld's proof of Theorem 3.2 for v non-archimedean see [BC].

If $\mathcal{N}^+$ is any ideal (or totally positive element) of F prime to the places of S, one can also define a Shimura curve $X_S(\mathcal{N}^+)$ by adding "auxiliary level structure" of level $\mathcal{N}^+$.

Denote by J_S and $J_S(\mathcal{N}^+)$ the jacobian varieties of X_S and $X_S(\mathcal{N}^+)$ respectively. The relevance of these jacobians is that they are expected to parametrise certain elliptic curves over F in the same way that jacobians of modular curves uniformise elliptic curves over $\mathbb{Q}$.

More precisely, a (modular, in the sense of Conjecture 3.1) elliptic curve E over F is said to be *arithmetically uniformisable* if there exists a Shimura curve $X_S(\mathcal{M})$ and a non-constant map of abelian varieties over F, generalising (4)

$$\Phi_{S,E} \colon J_S(\mathcal{M}) \longrightarrow E. \tag{15}$$

Conjecture 3.1 leads one to expect that many (*but not all*, in general!) elliptic curves over F are arithmetically uniformisable. More precisely,

Theorem 3.3. *A modular elliptic curve E over F is arithmetically uniformisable if and only if at least one of the following conditions holds.*

1. *The degree of F over $\mathbb{Q}$ is odd;*
2. *There is a place v of F for which* $\mathrm{ord}_v(\mathcal{N})$ *is odd.*

When condition 1 is satisfied, a Shimura curve uniformising E can be taken to be of the form $X_S(\mathcal{N}_E)$, where $S = S_\infty$ is the set of archimedean places of F. If condition 1 is not satisfied, but 2 is, one can consider a Shimura curve associated to $S = \{v\} \cup S_\infty$ with a suitable choice of level structure. See [Zh1] for more details on Shimura curves and their associated modular parametrisations.

3.2. Heegner points. From now on we assume that E/F is *semistable*, and that there is a factorisation $\mathcal{N} = \mathcal{N}^+\mathcal{N}^-$ of the conductor into a product of ideals with the property that the set of places of F

$$S := \{v \text{ divides } \infty \text{ or } \mathcal{N}^-\}$$

has odd cardinality. (Placing oneself in this special situation facilitates the exposition, and does not obscure any of the essential features we wish to discuss.) This assumption implies that E is arithmetically uniformisable and occurs as a quotient of the Jacobian $J_S(\mathcal{N}^+)$ of the Shimura curve $X_S(\mathcal{N}^+)$ of the previous section. Let

$$\Phi_{S,E}^{\mathcal{N}^+}\colon \operatorname{Div}^0(X_S(\mathcal{N}^+)) \longrightarrow E \tag{16}$$

denote the Shimura curve parametrisation attached to this data.

Just like classical modular curves, the curve $X_S(\mathcal{N}^+)$ is also equipped with a collection of CM points attached to certain CM extensions of F. More precisely, let K be a quadratic extension of F satisfying:

1. For all places $v \in S$, the F_v-algebra $K \otimes_v F_v$ is a field.
2. For all places $v|\mathcal{N}^+$, the F_v-algebra $K \otimes_v F_v$ is isomorphic to $F_v \oplus F_v$.

Note that condition 1 implies in particular that K is a CM extension of F, since S contains all the archimedean places of F.

Fix an $\mathcal{O}_F$-order $\mathcal{O}$ of K, and let H denote the associated ring class field of K. There is a canonical collection $CM(\mathcal{O}) \subset X_S(\mathcal{N}^+)(H)$ associated, in essence, to solutions to the moduli problem related to $X_S(\mathcal{N}^+)$ which have "extra endomorphisms by $\mathcal{O}$." This fact allows an extension of the theory of Heegner points to the context of totally real fields.

3.3. The efficient calculation of global points. Assume for notational simplicity that $\mathcal{N}^+ = 1$. From a computational perspective, it would be useful to have efficient numerical recipes for computing the points of $CM(\mathcal{O})$ and their images in $E(H)$ under the parametrisation $\Phi_{S,E}$ of (16). Difficulties arise in calculating Heegner points arising from Shimura curve parametrisations, largely because the absence of Fourier expansions for modular forms on $\Gamma\backslash\mathcal{H}_v$ prevents one from writing down an explicit analytic formula for $\Phi_{S,E}$ analogous to (6).

The article [El1] proposes to work with Shimura curves by computing algebraic equations for them. This approach can be carried out when the group Γ arising in an archimedean uniformisation of $X_S(\mathbb{C})$ following Theorem 3.2 is contained with small index in a Hecke triangle group. Adapting the ideas of [KM] to the context of Shimura curves might also yield a more systematic approach to these types of questions. Nonetheless, it appears that an approach relying on an explicit global equation for the Shimura curve may become cumbersome, since such an algebraic equation is expected to be quite complicated for even modest values of F and S.

Alternately, one may try to exploit the non-archimedean uniformisations of $X_S(\mathbb{C}_v)$ given by Theorem 3.2. Given a non-archimedean place $v \in S$, let B and R be the quaternion algebra and Eichler order associated to S and v as in the statement of this theorem. An F-algebra embedding

$$\Psi : K \longrightarrow B$$

is said to be *optimal* relative to $\mathcal{O}$ if $\Psi(K) \cap R = \Psi(\mathcal{O})$. It can be shown that the number of distinct optimal embeddings of K into B, up to conjugation by the normaliser of $R^\times$ in $B^\times$, is equal to the class number of $\mathcal{O}$. Let h denote this class number and let $\Psi_1, \ldots, \Psi_h$ be representatives for the distinct conjugacy classes of optimal embeddings of $\mathcal{O}$ into R. Let τ_j and $\bar{\tau}_j$ denote the fixed points for $\Psi_j(K^\times)$ acting on $\mathcal{H}_v$. Then the points in $CM(\mathcal{O})$ are identified with the points τ_j, $\bar{\tau}_j$ under the identification of Theorem 3.2.

In his thesis [Gre], Matthew Greenberg exploits this explicit v-adic description of the points in $CM(\mathcal{O})$ and computes their images in $E(\mathbb{C}_v)$ analytically. The absence of cusps on X_S and of the attendant Fourier expansion of modular forms is remedied in part by an alternate combinatorial structure on $X_S(\mathbb{C}_v)$ which allows explicit v-adic analytic calculations with cusp forms on X_S. This combinatorial structure arises from the *reduction map*

$$r : \mathcal{H}_v \longrightarrow \mathcal{T}$$

on $\mathcal{H}_v$, where $\mathcal{T}$ is the *Bruhat–Tits tree* of $\mathrm{PGL}_2(F_v)$, a homogeneous tree with valency $|v| + 1$. Thanks to this structure, rigid analytic modular forms of weight two on $\Gamma\backslash\mathcal{H}_v$ admit a simple description as functions on the edges of the quotient graph $\Gamma\backslash\mathcal{T}$ satisfying a suitable harmonicity property. (For a more detailed discussion of the description of rigid analytic modular forms on $\Gamma\backslash\mathcal{H}_v$ in terms of an associated Hecke eigenfunction on the edges of the Bruhat–Tits tree, see Chapters 5 and 6 of [Da2] for example.)

Greenberg explains how the knowledge of the eigenfunction on $\Gamma\backslash\mathcal{T}$ associated to E can be parlayed into an efficient algorithm for computing the Shimura curve parametrisation $\Phi_{S,E}$ of (16), viewed as a v-adic analytic map

$$\Phi_{S,E}^v : \mathrm{Div}^0(\Gamma\backslash\mathcal{H}_v) \longrightarrow E(\mathbb{C}_v).$$

The main ingredient in Greenberg's approach is the theory of "overconvergent modular symbols" developed in [PS], adapted to the context of automorphic forms on definite quaternion algebras.

For example, setting $\omega = \frac{1+\sqrt{5}}{2}$, Greenberg considers the elliptic curve

$$E : y^2 + xy + \omega y = x^3 + (-\omega - 1)x^2 + (-30\omega - 45)x + (-111\omega - 117)$$

defined over $F = \mathbb{Q}(\sqrt{5})$. This curve has conductor $\mathcal{N} = v = (3-5\omega)$, a prime ideal above 31. Consider the CM extension $K = F(\sqrt{-\omega - 5})$ of F. It has class number two, and its Hilbert class field is equal to $H = K(i)$ (where $i = \sqrt{-1}$) by genus

theory. Letting $\tau \in \mathcal{H}_v$ be an element of $CM(\mathcal{O}_K)$, and τ' its translate by the element of order 2 in the class group of K, Greenberg computes the image of $\Phi^v_{S,E}((\tau)-(\tau'))$ in $E(K_v)$ to a v-adic accuracy of 31^{-30}, obtaining a point that agrees with the global point

$$P = \left(\frac{578\omega - 1}{90}, -\frac{27178\omega + 9701}{2700} i - \frac{668\omega - 1}{180} \right)$$

to that degree of accuracy.

The calculations of [Gre] convincingly demonstrate that Heegner points arising from Shimura curve parametrisations can be computed fairly systematically in significant examples using the Cerednik–Drinfeld theory. It would be interesting to understand whether the archimedean uniformisations described in Theorem 3.2 can be similarly exploited.

3.4. Relation with L-series. Retaining the notations of the previous section, let P be any point of $CM(\mathcal{O}) \subset X_S(H)$, and let χ be a character of $G = \mathrm{Gal}(H/K)$. Suppose for simplicity that this character is non-trivial, so that

$$D_\chi := \sum_{\sigma \in G} \chi(\sigma) P^\sigma \text{ belongs to } \mathrm{Div}^0(X_s(H)) \otimes \mathbb{C}.$$

Let P_χ denote the image of D_χ,

$$P_\chi := \Phi_{S,E}(D_\chi).$$

The Heegner point P_χ enjoys the following property analogous to the formula of Gross and Zagier.

Theorem 3.4 (Zhang). *The height of P_χ is equal to an explicit non-zero multiple of $L'(E/K, \chi, 1)$.*

The proof of Theorem 3.4, which is explained in [Zh1], [Zh2], and [Zh3], proceeds along general lines that are similar to those of [GZ] needed to handle the case $F = \mathbb{Q}$, although significant new difficulties have to be overcome in handling Shimura curve parametrisations. Note that, even when $F = \mathbb{Q}$, Zhang's theorem asserts something new since an elliptic curve over $\mathbb{Q}$ may possess, along with the usual modular curve parametrisation, a number of Shimura curve parametrisations.

3.5. The Birch and Swinnerton-Dyer conjecture. Zhang's formula has applications to the arithmetic of elliptic curves defined over totally real fields that are analogous to those of the original Gross–Zagier formula.

Theorem 3.5. *Suppose that E is arithmetically uniformisable. Then conjectures BSD_0 and BSD_1 are true for E.*

Sketch of proof. Since E is arithmetically uniformisable, there is a Shimura curve $X_S(\mathcal{M})$ parametrising E, for an appropriate $\mathcal{M}|\mathcal{N}_E$. If $\mathrm{ord}_{s=1} L(E,s) \leq 1$, one can choose as in the proof of Theorem 2.11 an auxiliary quadratic CM extension K of F in which all the primes of S are inert, those dividing $\mathcal{M}$ are split, and for which

$$\mathrm{ord}_{s=1} L(E/K, s) = 1.$$

After choosing such a K, the Heegner point P_K attached to K and the parametrisation (15) is of infinite order by Theorem 3.4. A natural extension of Kolyvagin's Theorem 2.9 to the context of totally real fields has been proved by Kolyvagin and Logachev [KL]. Their result implies that P_K generates a subgroup of $E(K)$ of finite index, and that $Ш(E/K)$ is finite. Theorem 3.5 now follows much as in the proof of Theorem 2.11. □

The proof of Theorem 3.5 sketched above breaks down for elliptic curves that are not arithmetically uniformisable in the sense of Theorem 3.3. This is the case for the elliptic curve

$$E : y^2 + xy + \varepsilon^2 y = x^3, \quad \varepsilon = \frac{5+\sqrt{29}}{2} \in \mathcal{O}_F^\times. \tag{17}$$

defined over the real quadratic field $F = \mathbb{Q}(\sqrt{29})$ and having everywhere good reduction over F.

Remark 3.6. It should be noted however that the curve E of (17) is isogenous to a quotient of the modular Jacobian $J_1(29)$, this circumstance arising from the fact that E is a $\mathbb{Q}$-curve, i.e., is isogenous to its Galois conjugate. Hence a variant of the Heegner point construction exploiting CM points on $X_1(29)$ might provide some information on the arithmetic of E.

In light of this remark, an even more puzzling example is given by the following elliptic curve discovered by R. Pinch,

$$y^2 - xy - \omega y = x^3 + (2+2\omega)x^2 + (162+3\omega)x + (71+34\omega), \quad \omega = \frac{1+\sqrt{509}}{2}, \tag{18}$$

which has everywhere good reduction over $F = \mathbb{Q}(\sqrt{509})$, and is *not* isogenous to its Galois conjugate. The curve given by (18), and any of its quadratic twists over F, are elliptic curves for which no variant of the Heegner point construction relying on CM points is known. For such elliptic curves, the strategy of proof of Theorem 3.5 runs across a fundamental barrier.

In spite of this the following theorem has been proved independently in [Lo1], [Lo2] and [TZ].

Theorem 3.7 (Longo, Tian-Zhang). *Suppose that E is any (modular) elliptic curve over a totally real field F. Then conjecture BSD_0 is true for E.*

Sketch of proof. We indicate the idea of the proof in the simplest case where E has everywhere good reduction over a real quadratic field F. Let K be any CM extension of F, and fix a rational prime p. The key fact is that, even though E is *not* arithmetically uniformisable, it is still possible to produce a sequence $X_1, \ldots, X_n, \ldots$ of Shimura curves in such a way that the Galois module given by the p^n-torsion $E[p^n]$ of E appears as a Jordan–Hölder constituent of $J_n[p^n]$, where J_n denotes the Jacobian of X_n. The Shimura curve X_n is associated to the set $S_n := \{\ell_n, \infty_1, \infty_2\}$ of places of F, for a judiciously chosen (non-archimedean) place ℓ_n of F. The existence of X_n follows from the theory of congruences between modular forms and the Jacquet–Langlands correspondence. The Heegner point attached to K and X_n can then be used to produce, following a variant of Kolyvagin's original recipe, a global cohomology class in $H^1(K, J_n[p^n])$, and, from this, a class $\kappa_n \in H^1(K, E[p^n])$. A key formula, whose proof exploits the Cerednik–Drinfeld theory of ℓ_n-adic uniformisation of X_n, relates the restriction of κ_n in the local cohomology group $H^1(K_{\ell_n}, E[p^n])$ to the special value of $L(E/K, 1)$. (More precisely, to a suitable *algebraic part*, taken modulo p^n.) In particular, if this special value is non-zero, then the class κ_n is non-trivial for n sufficiently large. (In fact, this is even so locally at ℓ_n.) This local control of the classes κ_n is enough to prove (following the lines of Kolyvagin's original argument) that the p^n-Selmer group of E over K has cardinality bounded independently of n, and therefore that $E(K)$ and the p-primary component of $Ш(E/K)$ are both finite. The same finiteness results hold *a fortiori* with K replaced by F. It is in ensuring the existence of a suitable auxiliary CM field K for which $L(E/K, 1) \neq 0$ that the non-vanishing hypothesis on $L(E/F, 1)$ made in the statement of Conjecture BSD_0 is used in a crucial way. □

A similar approach to bounding the Selmer group of E relying on congruences between modular forms was first exploited in [BD3] where it was used to prove part of the "main conjecture" of Iwasawa Theory attached to an elliptic curve $E/\mathbb{Q}$ and the anticyclotomic $\mathbb{Z}_p$-extension of an imaginary quadratic field K.

Theorem 3.7 notwithstanding, the following question retains an alluring aura of mystery.

Question 3.8. Prove Conjecture BSD_1 for elliptic curves over totally real fields that are *not* arithmetically uniformisable.

For example, let E_0 be an elliptic curve with everywhere good reduction over a real quadratic field F such as the curve given in equations (17) and (18). Let K be a quadratic extension of F which is neither totally real nor complex, i.e., an extension with one complex and two real places. Let E denote the twist of E_0 by K. It can be shown that $\mathrm{sign}(E, F) = -1$, so that $L(E/F, s)$ vanishes to odd order. Can one show that $E(F)$ is infinite, if $L'(E/F, 1) \neq 0$? This would follow from a suitable variant of Theorem 2.5 or 3.4, but it is unclear how such a variant could be proved – or even formulated precisely! – in the absence of a known Heegner point construction for E.

4. Stark–Heegner points

Question 3.8 points out one among many instances where Heegner points are not sufficient to produce algebraic points on elliptic curves, even when the presence of such points is predicted by the Birch and Swinnerton-Dyer conjecture.

The notion of *Stark–Heegner point* is meant to provide a *conjectural* remedy by proposing constructions in a number of situations lying ostensibly outside the scope of the theory of complex multiplication.

4.1. ATR extensions of totally real fields. Let F be a totally real field of narrow class number 1, as in Section 3. A quadratic extension K of F is said to be *almost totally real* (or "ATR" for short) if it has exactly one complex place, so that the remaining real places split in K/F. The field K can be viewed as a subfield of $\mathbb{C}$ via its unique complex embedding. A point in the complex upper half-plane is called an *ATR point* if it generates an ATR extension of F. Let $\mathcal{H}'$ denote the set of all ATR points on $\mathcal{H}$, relative to a fixed real place v of F. Note that $\mathcal{H}'$ is preserved under the action of the Hecke congruence group $\Gamma_0(\mathcal{N}) \subset \mathrm{SL}_2(\mathcal{O}_F)$, although, because the action of this group is not discrete, the quotient $\Gamma_0(\mathcal{N})\backslash\mathcal{H}'$ inherits no obvious topology (other than the discrete one). Let f_E denote the Hilbert modular form of level $\mathcal{N}$ associated to E in Conjecture 3.1, and write

$$\omega_f := f_E(z_1, \ldots, z_v) dz_1 \ldots dz_v$$

for the corresponding $\Gamma_0(\mathcal{N})$-invariant differential form on $\mathcal{H}^v$. The article [DL] describes a kind of *natural substitute* of the modular parametrisation attached to E, denoted

$$\Phi_E^v \colon \Gamma_0(\mathcal{N})\backslash\mathcal{H}' \longrightarrow E(\mathbb{C}). \tag{19}$$

A precise description of this map is given in Chapter VIII of [Da2] as well as in [DL]. We will not recount the details of this construction here, mentioning only that Φ_E^v is defined in terms of the periods of ω_f. It is in that sense that it can be viewed as purely analytic, even though Φ_E^v does not extend to a holomorphic or even continuous map on $\mathcal{H}$ (as is apparent from the fact that $\Gamma_0(\mathcal{N})$ acts on $\mathcal{H}$ with dense orbits). We note that the definition of Φ_E^v is quite concrete and lends itself well to computer calculations. In fact, working with the Hilbert modular form attached to E has the added computational advantage that the fourier expansion of ω_f is available as an aid to computing its periods numerically.

The main conjecture that is spelled out precisely in [DL] is that the points $\{\Phi_E^v(\tau)\}_{\tau\in\mathcal{H}'\cap K}$ belong to ring class fields of the ATR extension K of F, and that they enjoy all the properties (Shimura reciprocity law, norm compatibility relations) of classical Heegner points. This conjecture is also tested numerically and used to produce global points on the elliptic curve of equation (17) in terms of periods of the associated Hilbert modular form over $\mathbb{Q}(\sqrt{29})$. A proof of the conjectures of [DL] (an admittedly tall order, at present) would presumably lead to a solution to Question 3.8 proceeding along much the same lines as the proof of Theorems 2.11 and 3.5.

4.2. Ring class fields of real quadratic fields. We return now to the setting where E is an elliptic curve over $\mathbb{Q}$. Little changes in the analysis of Remark 2.7 when the imaginary quadratic field is replaced by a *real* quadratic field. Hence, if K is such a field and $\mathrm{sign}(E, K) = -1$, one expects the presence of a systematic collection of points defined over various ring class fields of K. This is intriguing, since the theory of complex multiplication gives no means of producing these points.

Suppose now that the conductor of E is the form $N = pM$, where p is a prime that does not divide M, so that E has multiplicative reduction at p. Let K be a *real* quadratic field satisfying the following "modified" Heegner hypothesis:

1. All the primes dividing M are split in K;
2. The prime p is inert in K.

These conditions are analogous to the ones that are imposed in the setting of classical Heegner points, with the prime p now playing the role of ∞. It can be shown that $\mathrm{sign}(E, K) = -1$, and the same holds for all twists of $L(E/K, s)$ by ring class characters of conductor prime to N. The analysis carried out in Remark 2.7 therefore shows that if H is any ring class field of K of discriminant prime to N, one has the same inequality as in (12):

$$\mathrm{ord}_{s=1}(L(E/H, s) \geq [H : K]. \tag{20}$$

The article [Da1] describes a conjectural recipe for constructing certain canonical points in $E(H)$, which is expected to yield a subgroup of finite index in $E(H)$ whenever (20) is an equality.

The idea behind the construction of [Da1] is to attach p-adic periods to f in a way which formally suggests viewing f as a "mock Hilbert modular form" on $\Gamma\backslash(\mathcal{H}_p \times \mathcal{H})$, where $\Gamma \subset \mathrm{SL}_2(\mathbb{Z}[1/p])$ is the subgroup of matrices which are upper triangular modulo M. The construction of these p-adic periods, which is described in [Da1], is essentially elementary. The main ingredient that enters in their definition is the theory of *modular symbols* associated to f, which states that the period integral

$$I_f\{r\to s\} := \frac{1}{\Omega^+}\mathrm{Re}\left(\int_r^s 2\pi i f(z)\, dz\right), \quad r, s \in \mathbb{P}_1(\mathbb{Q}) \tag{21}$$

takes *integer values* for a suitable choice of "real period" $\Omega^+ \in \mathbb{R}$, which is, up to a non-zero rational multiple, the real period in the Néron lattice of E.

Further pursuing the analogy with the setting of Section 4.1, the counterpart of the set $\mathcal{H}'$ of ATR points in $\mathcal{H}$ (associated to the real quadratic base field F and a choice of real embedding) is the collection $\mathcal{H}'_p$ of elements of $\mathcal{H}_p$ which generate a *real quadratic* extension of $\mathbb{Q}$. In particular, after fixing a p-adic embedding $K \subset \mathbb{C}_p$, the set $\mathcal{H}'_p \cap K$ is *non-empty*. Mimicking the formal aspects of the definition of the map (19) of Section 4.1, with the complex periods attached to a Hilbert modular form replaced by the (p-adic) periods on $\mathcal{H}_p \times \mathcal{H}$ attached to f, leads to the definition of a "modular parametrisation" analogous to (19)

$$\Phi_E^p\colon \Gamma\backslash\mathcal{H}'_p \longrightarrow E(\mathbb{C}_p). \tag{22}$$

Let D be the discriminant of K, and choose a $\delta \in \mathbb{Z}[1/p]$ satisfying

$$\delta^2 \equiv D \pmod{M}.$$

Let $\mathcal{F}^D$ be the set of primitive binary quadratic forms $Ax^2 + Bxy + Cy^2$ with coefficients in $Z[1/p]$, satisfying

$$B^2 - 4AC = D, \quad M|A, \quad B \equiv \delta \pmod{M}.$$

(A quadratic form is said to be *primitive* in this context if the ideal of $\mathbb{Z}[1/p]$ generated by (A, B, C) is equal to $\mathbb{Z}[1/p]$.) The group Γ acts naturally on $\mathcal{F}^D$ by "change of variables", and the quotient $\Gamma\backslash\mathcal{F}^D$ is equipped with a natural simply transitive action of the class group G_D of K arising from the Gaussian composition law. (Or rather, the Picard group of $\mathcal{O}_K[1/p]$, but these coincide since p is inert in K.) This action is completely analogous to the action of G_D on $\Gamma_0(N0\backslash\mathcal{H}^D$ (for D a negative discriminant) that is described in Section 2.2. Choose an embedding of K into $\mathbb{C}_p$, and for each quadratic form $F = [A, B, C] \in \mathcal{F}^D$, let

$$\tau = \frac{-B + \sqrt{D}}{2A} \in \mathcal{H}_p \tag{23}$$

be the corresponding element of $\mathcal{H}_p$ satisfying $F(\tau, 1) = 0$. The set $\mathcal{H}_p^D$ of all τ that arise in this way is preserved under the action of Γ, and the natural assignment given by (23) induces a bijection

$$\Gamma\backslash\mathcal{F}^D \longrightarrow \Gamma\backslash\mathcal{H}_p^D.$$

Hence the target of this bijection inherits a simply transitive action of G_D. Denote this action by $(\sigma, \tau) \mapsto \tau^\sigma$, for all $\sigma \in G_D$ and $\tau \in \mathcal{H}_p^D$. Conjectures 5.9 and 5.15 of [Da1] predict that

Conjecture 4.1. 1. For all $\tau \in \mathcal{H}_p^D$, the point $\Phi_E^p(\tau)$ is defined over H.

2. If $\chi : G_D \longrightarrow \mathbb{C}^\times$ is a complex character, then the expression

$$\sum_{\sigma \in G_D} \chi(\sigma)\Phi_E^p(\tau^\sigma) \in E(H) \otimes \mathbb{C}$$

is non-zero if and only if $L'(E/K, \chi, 1) \neq 0$. In particular, the subgroup of $E(H)$ generated by the Stark–Heegner points $\Phi_E^p(\tau)$, as $\tau \in \Gamma\backslash\mathcal{H}_p^D$, has rank $h = [H : K]$ if and only if $\mathrm{ord}_{s=1} L(E/H, s) = h$.

A proof of Conjecture 4.1 would not yield any new information about Conjecture BSD_r for $r \leq 1$, since this conjecture is already known for elliptic curves over $\mathbb{Q}$. It would, however, give a proof of some new cases of the Birch and Swinnerton-Dyer conjecture for Mordell–Weil groups over ring class fields of real quadratic fields, following a simple extension of Kolyvagin's arguments which is explained in [BD1] and in Chapter X of [Da2]. See also [BDD] for other ways in which a strengthening of

Conjecture 4.1 to modular forms with non-rational Fourier coefficients would imply new cases of Conjecture BSD_0, by adapting the ideas that are used in the proof of Theorem 3.7.

Conjecture 4.1 has been extensively tested numerically in [DG]. A significant improvement of the algorithms of [DG], based on ideas of Pollack and Stevens which grew out of their theory of overconvergent modular symbols, as mentioned in Section 3.3, is described in [DP1]. These improvements make it possible to find global points of large height on E rather efficiently. For example, the Stark–Heegner point on the elliptic curve E of conductor 11 given by the equation

$$y^2 + y = x^3 - x^2 - 10x - 20$$

attached to the field $K = \mathbb{Q}(\sqrt{101})$ can be computed to an 11-adic accuracy of 11^{-100} in a few seconds on a standard computer. It can then be "recognized" as the global point in $E(\mathbb{Q}(\sqrt{101}))$ with x-coordinate equal to

$$x = \frac{1081624136644692539667084685116849}{24684654182277032144757997152010 0}.$$

Of course, the calculation of Stark–Heegner points also has applications to explicit class field theory for real quadratic fields analogous to those described in Section 2.4 for imaginary quadratic fields. For example, let E be the unique elliptic curve over $\mathbb{Q}$ of conductor $p = 79$, defined by the Weierstrass equation

$$y^2 + xy + y = x^3 + x^2 - 2x.$$

The prime p is inert in the real quadratic field $K = \mathbb{Q}(\sqrt{401})$, which has class number five. The 5 distinct representatives in $\mathcal{H}_{79}^{401}$ can be chosen to be

$$\tau = \frac{-19 + \sqrt{401}}{2}, \quad \frac{19 - \sqrt{401}}{4}, \quad \frac{-15 + \sqrt{401}}{8}, \quad \frac{17 - \sqrt{401}}{8}, \quad \frac{-17 + \sqrt{401}}{4}.$$

The x-coordinates of the corresponding Stark–Heegner points $\Phi_E^{79}(\tau)$ (computed modulo 79^{20}) appear to satisfy the polynomial

$$x^5 - 20x^4 + 47x^3 - 31x^2 + x + 3$$

whose splitting field is indeed the Hilbert class field H of K. In fact this calculation leads to the discovery of points in $E(H)$. The analogous polynomial, for the real quadratic field $\mathbb{Q}(\sqrt{577})$ of class number seven, is

$$x^7 - 22x^6 + 74x^5 - 51x^4 - 40x^3 + 32x^2 + 2x - 1.$$

These examples are chosen at random among the hundreds of calculations that were performed in [DP1] to test the conjectures of [Da1] numerically. (More such calculations could be performed by the interested reader using the publicly available software [DP2] for calculating Stark–Heegner points, written in Magma, which is documented in [DP1].)

4.3. Beyond totally real fields?. The assumption that E is defined over a totally real field F, although it arises naturally in considering automorphic forms and their associated Shimura varieties, is not a natural one from the point of view of the Diophantine study of elliptic curves. It would be just as desirable to understand elliptic curves defined over general number fields, and to have the means of tackling conjectures BSD_0 and BSD_1 for such curves.

The simplest case arises when E is an elliptic curve defined over an *imaginary* quadratic field, denoted by F (and not K as in Section 2, since now it plays the role of the "ground field" over which E is defined). Assume for simplicity that F has class number one, and let $\mathcal{N}$ denote the conductor of E.

As in Section 3, the Shimura–Taniyama conjecture predicts that E corresponds to an automorphic form f on $\mathrm{GL}_2(F)$, which gives rise, following the description given in [Cr1], to a differential form ω_f on the *hyperbolic upper half space*

$$\mathcal{H}^{(3)} := \mathbb{C} \times \mathbb{R}^{>0}.$$

This three-dimensional real manifold is equipped with a hyperbolic metric and an action of $\mathrm{SL}_2(\mathbb{C})$ by isometries, and the differential ω_f is *invariant* under the resulting action of the subgroup $\Gamma_0(\mathcal{N}) \subset SL_2(\mathcal{O}_F)$ consisting of matrices which are upper triangular modulo $\mathcal{N}$. Congruence subgroups of $\mathrm{SL}_2(\mathcal{O}_F)$ are examples of so-called *Bianchi groups*; for information about their structure and properties and further references, see [EGM] for example.

The modular form ω_f does not give rise to a modular parametrisation of E analogous to (4). In fact, the symmetric space $\mathcal{H}^{(3)}$ is not even endowed with a natural complex structure (since it has real dimension 3); therefore the quotient $\Gamma_0(\mathcal{N})\backslash\mathcal{H}^{(3)}$ cannot be viewed as the points of a complex analytic variety, much less an algebraic one. The absence of a Shimura variety attached to f implies that one has less control on the arithmetic of this modular form. For certain f, Taylor [Ta] has been able to construct the Galois representations which the Langlands conjectures attach to f by exploiting congruences with modular forms on GSp(4) whose associated Galois representations can be found in the cohomology of the appropriate Shimura varieties. Unfortunately, global points on elliptic curves or abelian varieties, unlike p-adic Galois representations (as in the work of Taylor) or Galois cohomology classes attached to rational points (as in the proof of Theorem 3.7), do not readily lend themselves to constructions based on congruences between modular forms.

Nonetheless, the differential form ω_f comes with an attendant notion of *modular symbol* which enjoys the same integrality properties as in the classical case. (For a discussion of modular symbols attached to forms on $\mathrm{GL}_2(F)$, and their computational applications, see [Cr1], [CW].) Trifkovic [Tr] exploits this modular symbol to transpose to f the definition of the p-adic periods on $\mathcal{H}_p \times \mathcal{H}$ alluded to in Section 4.2. In this way he associates to ω_f a "modular form on $\Gamma\backslash(\mathcal{H}_\mathfrak{p} \times \mathcal{H}^{(3)})$", where

1. $\mathfrak{p}$ is a prime of K dividing $\mathcal{N} = \mathcal{M}\mathfrak{p}$ exactly;

2. $\mathcal{H}_\mathfrak{p} = \mathbb{P}_1(\mathbb{C}_p) - \mathbb{P}_1(F_\mathfrak{p})$ is the $\mathfrak{p}$-adic upper half plane (defined after choosing an embedding $F_\mathfrak{p} \subset \mathbb{C}_p$);

3. $\Gamma \subset \mathrm{SL}_2(\mathcal{O}_F[1/\mathfrak{p}])$ is the subgroup consisting of matrices which are upper triangular modulo $\mathcal{M}$.

The set $\mathcal{H}'_\mathfrak{p}$ is simply the set of $\tau \in \mathcal{H}_\mathfrak{p}$ which generate a quadratic extension of $F \subset F_\mathfrak{p}$. Trifkovic uses his $\mathfrak{p}$-adic periods to define an explicit, numerically computable map

$$\Phi_E^\mathfrak{p} \colon \Gamma\backslash\mathcal{H}'_\mathfrak{p} \longrightarrow E(\mathbb{C}_p),$$

and formulates an analogue of Conjecture 4.1 for this map, predicting that $\Phi_E^\mathfrak{p}(\tau)$ is defined over a specific ring class field of $K = F(\tau)$ for all $\tau \in \mathcal{H}'_p$.

Trifkovic has been able to gather extensive numerical evidence for his "Stark–Heegner conjectures" in this setting. Here is just one example taken among the many calculations that are reported on in [Tr]. Let E be the elliptic curve over $F = \mathbb{Q}(\sqrt{-11})$ given by the Weierstrass equation

$$y^2 + y = x^3 + \left(\frac{1-\sqrt{-11}}{2}\right)x^2 - x.$$

Its conductor is the prime $\mathfrak{p} = 6+\sqrt{-11}$ of F of norm 47. Note that E is not isogenous to its Galois conjugate, since $\mathfrak{p} \neq \bar{\mathfrak{p}}$. The quadratic extension $K = F(\sqrt{29})$ has class number 5. Trifkovic computes the five distinct Stark–Heegner points attached to the maximal order of K, as elements of $E(\mathbb{Q}_{47})$ (using the isomorphism $K_\mathfrak{p} = \mathbb{Q}_{47}$) with an accuracy of 20 significant 47-adic digits. This allows him to "guess" that the x coordinates of these Stark–Heegner points satisfy the degree 5 polynomial

$$x^5 - \left(\frac{80299+139763\sqrt{-11}}{149058}\right)x^4 + \left(\frac{-558203+71567\sqrt{-11}}{149058}\right)x^3 + \left(\frac{141709+45575\sqrt{-11}}{74529}\right)x^2 + \left(\frac{8372-7727\sqrt{-11}}{24843}\right)x + \left(\frac{-473+35\sqrt{-11}}{2366}\right)$$

whose splitting field can then be checked to be the Hilbert class field of K.

For many more calculations of this type, and a precise statement of the conjecture on which they rest, the reader is invited to consult [Tr].

4.4. Theoretical evidence. In spite of the convincing numerical evidence that has been gathered in their support, the conjectures on Stark–Heegner points suffer from the same paucity of theoretical evidence as in the setting of Stark's original conjectures on units. What little evidence there is at present can be grouped roughly under the following two rubrics:

4.4.1. Stark–Heegner points and Stark units. Many of the basic theorems and applications of elliptic curves have counterparts for units of number fields. (For instance, the Mordell–Weil theorem is analogous to Dirichlet's Unit Theorem; Lenstra's factorisation algorithm based on elliptic curves, to the Pollard $p-1$ method; to name just two examples.) The very terminology "Stark–Heegner point" is intended to convey the idea that these points are analogous to Stark units constructed from special values of L-series.

To make this sentiment precise, one can replace the cusp forms that enter into the constructions of Section 4.2 by *modular units*, or rather, their logarithmic derivatives which are Eisenstein series of weight two. Pursuing this idea, the article [DD] associates to any modular unit α on $\Gamma_0(N)\backslash\mathcal{H}$ and to $\tau \in \mathcal{H}_p' \cap K$ where K is a real quadratic field in which p is inert, an element $u(\alpha, \tau) \in K_p^\times$, which is predicted to behave like an elliptic unit defined over a ring class field of an imaginary quadratic field. More precisely, if $\mathcal{O}$ is the order associated to τ, and H denotes the corresponding ring class field of K, it is conjectured that $u(\alpha, \tau)$ belongs to $\mathcal{O}_H[1/p]^\times$ and obeys a Shimura reciprocity law formulated exactly as in Conjecture 4.1.

Section 3.1 [DD] attaches to α and to τ a ζ-function $\zeta(\alpha, \tau, s)$ which is essentially (up to a finite collection of Euler factors depending on α) the partial zeta-function attached to K and the narrow ideal class corresponding to τ. In particular, this zeta-function has a meromorphic continuation to $\mathbb{C}$ with at worst a simple pole at $s = 1$. Sections 4.1-4.3 of [DD] explain how a p-adic zeta function $\zeta_p(\alpha, \tau, s)$ can be defined by p-adically interpolating the special values of $\zeta(\alpha, \tau, k)$ at certain negative integers.

The main result of [DD], which is contained in Theorems 3.1 and 4.1 of [DD], is then

Theorem 4.2. *For all $\tau \in \mathcal{H}_p'$,*

$$\zeta(\alpha, \tau, 0) = \frac{1}{12}\,\mathrm{ord}_p(u(\alpha, \tau)); \tag{24}$$

$$\zeta_p'(\alpha, \tau, 0) = -\frac{1}{12}\log_p \mathrm{Norm}_{K_p/\mathbb{Q}_p}(u(\alpha, \tau)). \tag{25}$$

This theorem is consistent with Gross's p-adic analogue of the Stark conjectures [Gr1], [Gr2], which expresses the left hand side of (25) in terms of p-adic logarithms of the norm to $\mathbb{Q}_p$ certain global p-units in abelian extensions of K. We note that the conjecture of [DD] represents a genuine *refinement* of Gross's conjecture in the special case of ring class fields of real quadratic fields, since it gives a formula for the Gross–Stark units *as elements of* $K_p^\times$, and not just for their norms to $\mathbb{Q}_p^\times$.

Remark 4.3. A purely archimedean analogue of the setting of Theorem 4.2 is considered in [CD], leading to the conjectural construction of units in abelian extensions of an ATR extension K of a totally real field F in terms of periods of weight two Eisenstein series on the Hilbert modular group attached to F. This construction can be viewed either as the archimedean analogue of the main construction of [DD], or

as the analogue of the main construction of [DL] in which cusp forms on the Hilbert modular group are replaced by Eisenstein series. This construction (in the setting of abelian extensions of K) goes further than the original Stark conjectures by proposing a formula for the Stark units as elements of $\mathbb{C}^\times$, and not just for their *lengths* which are expressed in terms of values of L-series at $s = 0$. In other words, the formulae of [CD] capture the *arguments* as well as the absolute values of these Stark units (relative to a complex embedding of the ring class field H of K extending the unique complex embedding of K.)

Remark 4.4. The proof of Theorem 4.2 brings to light the role of the Eisenstein series of weight k and their associated periods (with k a weight which can be taken to vary p-adically) in relating the invariants $u(\alpha, \tau)$ to special values of L-functions. This suggests that the Stark–Heegner points of Section 4.2 should be related to the periods of a *Hida family* interpolating the cuspidal eigenform f in weight two.

4.4.2. The rationality of Stark–Heegner points over genus fields. Returning to the setting of Section 4.2, let K be a real quadratic field of discriminant D satisfying the auxiliary hypotheses relative to N that were mentioned in Section 4.2, and let H be its Hilbert class field. Write $G_D = \mathrm{Gal}(H/K)$ as before.

To each factorisation $D = D_1 D_2$ of D as a product of two fundamental discriminants is associated the unramified quadratic extension $L = \mathbb{Q}(\sqrt{D_1}, \sqrt{D_2}) \subset H$ of K. This field corresponds to a quadratic character

$$\chi : G_D \longrightarrow \pm 1,$$

called the *genus character* associated to the factorisation (D_1, D_2). Let χ_1 and χ_2 denote the quadratic Dirichlet characters attached to $\mathbb{Q}(\sqrt{D_1})$ and $\mathbb{Q}(\sqrt{D_2})$ respectively. Then χ, viewed as a character of the ideals of K, is characterised on ideals prime to D by the rule

$$\chi(\mathfrak{n}) = \chi_1(\mathrm{Norm}\,\mathfrak{n}) = \chi_2(\mathrm{Norm}\,\mathfrak{n}).$$

The field L is also called the *genus field* of K attached to (D_1, D_2). Let $E(L)^\chi$ denote the submodule of the Mordell–Weil group $E(L)$ on which G_D acts via the character χ.

Recall the action of G_D on $\Gamma\backslash\mathcal{H}_p^D$ arising from its identification with the class group of K. Define the point

$$P_\chi = \sum_{\sigma \in G_D} \chi(\sigma) \Phi_E^p(\tau^\sigma) \in E(K_p).$$

Conjecture 4.1 predicts that this local point belongs to $E(L)^\chi$ (after fixing an embedding $L \subset K_p$), and that it is of infinite order if and only if

$$L(E/K, \chi, s) = L(E, \chi_1, s) L(E, \chi_2, s)$$

has a simple zero at $s = 1$.

For each $m|N$ with $\gcd(m, N/m) = 1$, let w_m denote the sign of the Fricke involution at m acting on f. Note that the modified Heegner hypothesis implies that $\chi_1(-M) = \chi_2(-M)$. The main result of [BD5] is

Theorem 4.5. *Suppose that E has at least two primes of multiplicative reduction, and that $\chi_1(-M) = -w_M$.*

1. *There is a global point $\boldsymbol{P}_\chi \in E(L)^\chi$ and $t \in \mathbb{Q}^\times$ such that*

$$P_\chi = t\,\boldsymbol{P}_\chi \quad \text{in } E(K_p) \otimes \mathbb{Q}. \tag{26}$$

2. *The point $\boldsymbol{P}_\chi$ is of infinite order if and only if $L'(E/K, \chi, 1) \neq 0$.*

The proof of Theorem 4.5 relies on the connection between Stark–Heegner points and Shintani-type periods attached to Hida families alluded to in Remark 4.4, which grew out of the calculations of [DD]. A second key ingredient is the relation, made precise in [BD2] and [BD4], between classical Heegner points arising from certain Shimura curve parametrisations and the derivatives of associated two-variable anticyclotomic p-adic L-functions attached to Hida families. In a nutshell, these two ingredients are combined to express the Stark–Heegner point P_χ as a classical Heegner point, following an idea whose origins (as is explained in the introduction of [BD5]) can be traced back to Kronecker's "solution of Pell's equation" in terms of special values of the Dedekind eta-function.

Remark 4.6. A result of Gross–Kohnen–Zagier [GKZ] suggests that the position of the Stark–Heegner point P_χ in the Mordell–Weil group $E(L)^\chi$ is controlled by the Fourier coefficients of a modular form of weight $3/2$ associated to f via the Shimura lift. See [DT] where results of this type are discussed.

References

[BC] Boutot, J.-F., Carayol, H., Uniformisation p-adique des courbes de Shimura: les théorèmes de Čerednik et de Drinfeld. in *Courbes modulaires et courbes de Shimura* (Orsay, 1987/1988), *Astérisque* **196–197** (1991), 45–158.

[BCDT] Breuil, C., Conrad, B., Diamond, F., Taylor, R., On the modularity of elliptic curves over **Q**: wild 3-adic exercises. *J. Amer. Math. Soc.* **14** (2001), 843–939.

[BCdeSGKK] Bump, D., Cogdell, J. W., de Shalit, E., Gaitsgory, D., Kowalski, E., Kudla, S. S., *An introduction to the Langlands program*. Lectures presented at the Hebrew University of Jerusalem (ed. by J. Bernstein and S. Gelbart), Birkhäuser, Boston, MA, 2003.

[BD1] Bertolini, M., Darmon, H., Kolyvagin's descent and Mordell-Weil groups over ring class fields. *J. Reine Angew. Math.* **412** (1990), 63–74.

[BD2] Bertolini, M., Darmon, H., Heegner points, p-adic L-functions, and the Cerednik-Drinfeld uniformization. *Invent. Math.* **131** (3) (1998), 453–491.

[BD3] Bertolini, M., Darmon, H., Iwasawa's Main Conjecture for Elliptic Curves over Anticyclotomic $\mathbb{Z}_p$-extensions, *Ann. of Math.* **162** (2005), 1–64.

[BD4] Bertolini, M., Darmon, H., Hida families and rational points on elliptic curves. Submitted.

[BD5] Bertolini, M., Darmon, H., The rationality of Stark-Heegner points over genus fields of real quadratic fields. Submitted.

[BDD] Bertolini, M., Darmon, H., Dasgupta, S., Stark-Heegner points and special values of L-series. In *Proceedings of the Durham Symposium on L-functions and Galois representations* (July 2004), to appear.

[BDG] Bertolini, M., Darmon, H., Green, P., Periods and points attached to quadratic algebras. In *Heegner points and Rankin L-series* (ed. by H. Darmon and S. Zhang), Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 323–367.

[Bi] Birch, B., Heegner points: the beginnings. In *Heegner points and Rankin L-series* (ed. by H. Darmon and S. Zhang), Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 1–10.

[BS] Birch, B., Stephens, N., Computation of Heegner points. In *Modular forms* (Durham, 1983), Ellis Horwood Ser. Math. Appl.: Statist. Oper. Res., Horwood, Chichester 1984, 13–41.

[CD] Charollois, P., Darmon, H., Périodes des séries d'Eisenstein et arguments des unités de Stark. In progress.

[Cr1] Cremona, J. E., Hyperbolic tessellations, modular symbols, and elliptic curves over complex quadratic fields. *Compositio Math.* **51** (3) (1984), 275–324.

[Cr2] Cremona, J. E., *Algorithms for modular elliptic curves*. Second edition. Cambridge University Press, Cambridge 1997.

[CW] Cremona, J. E., Whitley, E. Periods of cusp forms and elliptic curves over imaginary quadratic fields. *Math. Comp.* **62** (205) (1994), 407–429.

[Da1] Darmon, H., Integration on $\mathcal{H}_p \times \mathcal{H}$ and arithmetic applications. *Ann. of Math.* (2) **154** (3) (2001), 589–639.

[Da2] Darmon, H., *Rational points on modular elliptic curves.* CBMS Reg. Conf. Ser. Math. 101, Amer. Math. Soc., Providence, RI, 2004.

[DD] Darmon, H., Dasgupta, S., Elliptic units for real quadratic fields. *Ann. of Math.* **163** (2006), 301–346.

[DG] Darmon, H., Green, P., Elliptic curves and class fields of real quadratic fields: algorithms and evidence. *Experiment. Math.* **11** (1) (2002), 37–55.

[DL] Darmon, H., Logan, A., Periods of Hilbert modular forms and rational points on elliptic curves. *Internat. Math. Res. Notices* **40** (2003), 2153–2180.

[DP1] Darmon, H., Pollack, R., Efficient calculation of Stark-Heegner points via overconvergent modular symbols. *Israel J. Math.* **153** (2006), 319–354.

[DP2] Darmon, H., Pollack, R., The `shp` package, Software written in Magma. Downloadable from http://www.math.mcgill.ca/darmon/programs/shp/shp.html.

[DT] Darmon, H., Tornaria, G., A Gross-Kohnen Zagier theorem for Stark-Heegner points. In progress.

[El1] Elkies, N. D., Shimura curve computations. In *Algorithmic number theory* (Portland, OR, 1998), Lecture Notes in Comput. Sci. 1423, Springer-Verlag, Berlin 1998, 1–47.

[El2] Elkies, N. D., Heegner point computations. In *Algorithmic number theory* (Ithaca, NY, 1994) Lecture Notes in Comput. Sci. 877, Springer-Verlag, Berlin 1994, 122–133.

[EGM] Elstrodt, J., Grunewald, F., Mennicke, J., *Groups acting on hyperbolic space*. Springer Monogr. Math., Springer-Verlag, Berlin 1998.

[Fu] Fujiwara, K., Modular varieties and Iwasawa theory. In *Algebraic number theory and related topics* (Kyoto, 1996), Sūrikaisekikenkyūsho Kūokyūroku (RIMS, Kyoto) **998**, (1997), 1–19.

[Ge] Gelbart, S. S., *Automorphic forms on adèle groups.* Ann. of Math. Stud. 83, Princeton University Press/University of Tokyo Press, Princeton, N.J./Tokyo 1975.

[GKZ] Gross, B., Kohnen, W., Zagier, D., Heegner points and derivatives of L-series. II. *Math. Ann.* **278** (1–4) (1987), 497–562.

[Go] Goldfeld, D., The Gauss class number problem for imaginary quadratic fields. In *Heegner points and Rankin L-series* (ed. by H. Darmon and S. Zhang), Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge, 2004, 25–36.

[Gre] Greenberg, M., Heegner points and rigid analytic modular forms. PhD thesis, McGill University, 2006.

[Gr1] Gross, B. H., p-adic L-series at $s = 0$. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **28** (3) (1981), 979–994.

[Gr2] Gross, B. H., On the values of abelian L-functions at $s = 0$. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **35** (1) (1988), 177–197.

[Gr3] Gross, B. H., Kolyvagin's work on modular elliptic curves. In *L-functions and arithmetic* (Durham, 1989), London Math. Soc. Lecture Note Ser. 153, Cambridge University Press, Cambridge 1991, 235–256.

[Gr4] Gross, B. H., Heegner points and representation theory. In *Heegner points and Rankin L-series* (ed. by H. Darmon and S. Zhang), Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 37–65.

[GZ] Gross, B. H., Zagier, D. B., Heegner points and derivatives of L-series. *Invent. Math.* **84** (2) (1986), 225–320.

[Ki] Kisin, M., Geometric deformations of modular Galois representations. *Invent. Math.* **157** (2) (2004), 275–328.

[KL] Kolyvagin, V. A., Logachev, D. Yu., Finiteness of $Ш$ over totally real fields. *Izv. Akad. Nauk SSSR Ser. Mat.* **55** (4) (1991), 851–876; English transl. *Math. USSR-Izv.* **39** (1) (1992), 829–853.

[KM] Khuri-Makdisi, K., Moduli interpretation of Eisenstein series. In progress.

[Ko] Kolyvagin, V. A., Finiteness of $E(\mathbb{Q})$ and $Ш(E, \mathbb{Q})$ for a subclass of Weil curves. *Izv. Akad. Nauk SSSR Ser. Mat.* **52** (3) (1988), 522–540, 670–671; English transl. *Math. USSR-Izv.* **32** (3) (1989), 523–541.

[Le] Lenstra, H. W., Solving the Pell Equation. *Notices Amer. Math. Soc.* **49** (2002), 182–192.

[Lo1] Longo, M., On the Birch and Swinnerton-Dyer conjecture over totally real fields. PhD thesis, University of Padova, 2004.

[Lo2] Longo, M., On the Birch and Swinnerton-Dyer conjecture for modular elliptic curves over totally real fields. *Ann. Inst. Fourier*, to appear.

[Ma] Mazur, B., Modular curves and the Eisenstein ideal. *Inst. Hautes Études Sci. Publ. Math.* **47** (1977), 33–186.

[Men] Mennicke, J., On Ihara's modular group. *Invent. Math.* **4** (1967), 202–228.

[Mer] Merel, L., Bornes pour la torsion des courbes elliptiques sur les corps de nombres. *Invent. Math.* **124** (1–3) (1996), 437–449.

[MM] Murty, M. R., Murty, V. K., *Non-vanishing of L-functions and applications*. Progr. Math. 157, Birkhäuser, Basel 1997.

[PS] Pollack, R., Stevens, G., Explicit computations with overconvergent modular symbols. In preparation.

[RV] Ricotta, G., Vidick, T., Hauteur asymptotique des points de Heegner. *Canad. J. Math.*, to appear.

[Se] Serre, J.-P., Complex multiplication. In *Algebraic Number Theory* (Brighton, 1965), Thompson, Washington D.C. 1967, 292–296.

[SW] Skinner, C. M., Wiles, A. J., Residually reducible representations and modular forms. *Inst. Hautes Études Sci. Publ. Math.* No. **89** (1999), 5–126.

[Ta] Taylor, R., l-adic representations associated to modular forms over imaginary quadratic fields. II. *Invent. Math.* **116** (1–3) (1994), 619–643.

[Tr] Trifkovic, M., Stark-Heegner points on elliptic curves over imaginary quadratic fields. Submitted.

[TZ] Tian, Y., Zhang, S., Book project in progress.

[Zh1] Zhang, S., Heights of Heegner points on Shimura curves. *Ann. of Math.* (2) **153** (2001), 27–147.

[Zh2] Zhang, S., Gross-Zagier formula for GL_2. *Asian J. Math.* **5** (2) (2001), 183–290.

[Zh3] Zhang, S., Gross-Zagier formula for GL(2). II. In *Heegner points and Rankin L-series* (ed. by H. Darmon and S. Zhang), Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 191–214.

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, QC, Canada

E-mail: darmon@math.mcgill.ca

Galois deformations and arithmetic geometry of Shimura varieties

Kazuhiro Fujiwara*

Abstract. Shimura varieties are arithmetic quotients of locally symmetric spaces which are canonically defined over number fields. In this article, we discuss recent developments on the reciprocity law realized on cohomology groups of Shimura varieties which relate Galois representations and automorphic representations.

Focus is put on the control of ℓ-adic families of Galois representations by ℓ-adic families of automorphic representations. Arithmetic geometrical ideas and methods on Shimura varieties are used for this purpose. A geometrical realization of the Jacquet–Langlands correspondence is discussed as an example.

Mathematics Subject Classification (2000). 11F55, 11F80, 11G18.

Keywords. Galois representations, automorphic representations, Shimura variety.

1. Introduction

For a reductive group G over $\mathbb{Q}$ and a homogeneous space X under $G(\mathbb{R})$ (we assume that the stabilizer K_∞ is a maximal compact subgroup modulo center), the associated modular variety is defined by

$$M_K(G, X) = G(\mathbb{Q})\backslash G(\mathbb{A}_f) \times X/K.$$

Here $\mathbb{A}_f = \prod'_{p \text{ prime}} \mathbb{Q}_p$ (the restricted product) is the ring of finite adeles of $\mathbb{Q}$, and K is a compact open subgroup of $G(\mathbb{A}_f)$.

$M_K(G, X)$ has finitely many connected components, and any connected component is an arithmetic quotient of a Riemannian symmetric space. It is very important to view $\{M_K\}_{K \subset G(\mathbb{A}_f)}$ as a projective system as K varies, and hence as a tower of varieties. The projective limit

$$M(G, X) = \varprojlim_{K \subset G(\mathbb{A}_f)} M_K(G, X)$$

admits a right $G(\mathbb{A}_f)$-action, which reveals a hidden symmetry of this tower. At each finite level, the symmetry gives rise to Hecke correspondences. For two compact open

*The author was partially supported by the 21st century COE program of JSPS. The author is grateful to the Research Institute for Mathematical Sciences of Kyoto University, and Columbia University for hospitality during the writing of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

subgroups $K, K' \subset G(\mathbb{A}_f)$ and $g \in G(\mathbb{A}_f)$, g defines a correspondence $[K'gK]$: the first projection $M_{K\cap g^{-1}K'g} \to M_K$, and the second is the projection $M_{K\cap g^{-1}K'g} \to M_{g^{-1}K'g} \simeq M_{K'}$.

The Betti cohomology group $H^*_B(M_K(G, X), \mathbb{Q})$ admits the action of the Hecke algebra $H(G(\mathbb{A}_f), K)_{\mathbb{Q}}$, the convolution algebra formed by the compactly supported $\mathbb{Q}$-valued K-biinvariant functions on $G(\mathbb{A}_f)$. The characteristic function χ_{KgK} of a double coset KgK acts as correspondence $[Kg^{-1}K]$.

When all connected components of X are hermitian symmetric spaces, we call $M_K(G, X)$ a Shimura variety and denote it by $\mathrm{Sh}_K(G, X)$. The arithmetic theory was begun by Shimura, and developed in his series of articles (cf. [44], [45], [47], [48]).

Shimura varieties have the following special features[1]:

- They are quasi-projective algebraic varieties over $\mathbb{C}$ (and smooth when K is small enough).
- They are canonically defined over the number field $E(G, X)$ (Shimura's theory of canonical models). $E(G, X)$ is called the reflex field. For a field extension E' of E, $\mathrm{Sh}(G, X)_{E'}$ denotes this model as defined over E'.

Shimura established the existence of canonical models when they are moduli spaces of abelian varieties with PEL structure [48], and even in some cases which are not moduli spaces of motives (exotic models, see [47])[2]. The functoriality of the canonical models is Shimura's answer to Hilbert's 12th problem (explicit construction of class fields).

In the case of Shimura varieties, $\mathrm{Sh}(G, X)_E = \varprojlim_{K\subset G(\mathbb{A}_f)} \mathrm{Sh}_K(G, X)_E$ is canonically defined over $E = E(G, X)$, which yields more symmetries than other modular varieties. The Galois group $G_E = \mathrm{Gal}(\bar{E}/E)$ acts on the tower[3], and hence the symmetry group is enlarged to $G_E \times G(\mathbb{A}_f)$. This symmetry acts on ℓ-adic étale cohomology groups $H^*_{\text{ét}}$ instead of the Betti cohomology H^*_B. The decomposition of the cohomology groups as Galois–Hecke bimodules is *the non-abelian reciprocity law* for Shimura varieties.

Aside from Shimura varieties, let us mention one more case which plays a very important role in the p-adic study of automorphic forms. When G is $\mathbb{Q}$-compact and X is a point, Hida noticed the arithmetic importance of $M(G, X)$ [21], though it is not a Shimura variety in general. These are called Hida varieties. Hida varieties are zero-dimensional, but the action of $G(\mathbb{A}_f)$ is quite non-trivial.

2. Non-abelian class field theory

We make the reciprocity law more explicit. Let $q(G)$ be the complex dimension of $\mathrm{Sh}(G, X)$. The ℓ-adic intersection cohomology group $IH^*_{\text{ét}}(\mathrm{Sh}_K(G, X)_{\bar{E}}) =$

[1] We assume that (G, X) satisfies Deligne's axioms [11], [12] for non-connected Shimura varieties.

[2] The general solution is due to Borovoi and Milne.

[3] In the following, for a field F, the absolute Galois group $\mathrm{Gal}(\bar{F}/F)$ of F is denoted by G_F.

$IH^*_{\text{ét}}(\overline{\mathrm{Sh}(G, X)}_{\min,\overline{E}}, \overline{\mathbb{Q}}_\ell)$ of the minimal compactification $\overline{\mathrm{Sh}(G, X)}_{\min,E}$ is pure in each degree, and the most interesting part lies in the middle degree $IH^{q(G)}_{\text{ét}}$. In general, the decomposition of IH^* is understood by Arthur packets, by granting Arthur's celebrated conjectures on local and global representations of reductive groups[4]. See [32] for a conjectural description in the general case.

Let us take a simple (but still deep) example. For a totally real number field F of degree g, let $I_F = \{\iota\colon F \hookrightarrow \mathbb{R}\}$ be the set of all real embeddings (regarded as the set of infinite places of F). One chooses a quaternion algebra D central over F which is split at one $\iota_0 \in I_F$ and ramifies at the other infinite places.

$G_D = \mathrm{Res}_{F/\mathbb{Q}} D^\times$ is the Weil restriction of the multiplicative group of D, and $X_D = \mathbb{P}^1_{\mathbb{C}} \backslash \mathbb{P}^1_{\mathbb{R}} = \mathbb{C}\backslash\mathbb{R}$ is the Poincaré double half plane. The resulting $\mathrm{Sh}(G_D, X_D)$ is a system of algebraic curves, namely the modular curve for $D = M_2(\mathbb{Q})$ and Shimura curves for the other cases, which has a canonical model over $F = E(G_D, X_D)$.

We fix a field isomorphism $\overline{\mathbb{Q}}_\ell \simeq \mathbb{C}$. The decomposition of $IH^1_{\text{ét}}$ as a $G_F \times H(G(\mathbb{A}_f), K)$-module is given by

$$IH^1_{\text{ét}}(\mathrm{Sh}_K(G_D, X_D)_{\overline{F}}) = \oplus_{\pi\in\mathcal{A}_D} \rho_\pi \otimes \pi_f^K.$$

Here $\mathcal{A}_D$ is the set of irreducible cuspidal representations $\pi = \pi_f \otimes \pi_\infty$ of $G_D(\mathbb{A})$ such that the infinite part π_∞ is cohomological for the trivial coefficient and the Jacquet–Langlands correspondent $JL(\pi)$ on $\mathrm{GL}_2(\mathbb{A}_F)$ is cuspidal[5]. $\rho_\pi\colon G_F \to \mathrm{GL}_2(\overline{\mathbb{Q}}_\ell)$ is the ℓ-adic Galois representation attached to π which has the same L-function as π. The identity

$$L(s, \rho_\pi) = L(s, \pi, \mathrm{st})$$

holds, where $L(s, \rho_\pi)$ is the Artin–Hasse–Weil L-function of ρ_π, and $L(s, \pi, \mathrm{st})$ is the standard Hecke L-function of π. This identity (or rather identities between their local factors) is the *non-abelian reciprocity law* realized on $IH^1_{\text{ét}}$. The modular curve case is due to Eichler and Shimura. The Shimura curve case is due to Shimura [47], Ohta [35] and Carayol [5]. It fits into the general representation theoretical framework due to Langlands. The Langlands correspondence is a standard form of non-abelian class field theory which predicts a correspondence between Galois and automorphic representations.

In the following, we discuss the relationship between ℓ-adic families of Galois representations and ℓ-adic families of automorphic representations (with applications to the Langlands correspondence as Wiles did for elliptic curves over $\mathbb{Q}$ in his fundamental work [57]). We show how arithmetic geometrical ideas and methods on Shimura varieties are effectively used for this purpose[6].

[4]It is more natural (and sometimes more convenient because of vanishing theorems known in harmonic analysis) to introduce coefficient sheaves attached to representations of G.

[5]The finite part π_f of π is defined over $\overline{\mathbb{Q}}_\ell$ via identification $\overline{\mathbb{Q}}_\ell \simeq \mathbb{C}$.

[6]Though the viewpoint of motives is extremely important, an emphasis is put on Galois representations in this article.

3. Galois deformations and nearly ordinary Hecke algebras for GL_2

In this section F denotes a totally real field, and I_F is as in §2. Fix an ℓ-adic field E_λ with integer ring o_λ and the maximal ideal λ.

By $I_{F,\ell}$ we mean the set of all field embeddings $F \hookrightarrow \overline{\mathbb{Q}}_\ell$. $I_{F,\ell}$ is canonically identified with $\coprod_{v|\ell} I_{F_v}$, where F_v is the local field at v, and I_{F_v} is the set of all continuous embeddings $F_v \hookrightarrow \overline{\mathbb{Q}}_\ell$. $I_{F,\ell}$ is identified with I_F by the isomorphism $\overline{\mathbb{Q}}_\ell \simeq \mathbb{C}$ chosen in §2.

Let Σ be a set of finite places which contains all places dividing ℓ. Let G_Σ denote $\mathrm{Gal}(F_\Sigma/F)$, where F_Σ is the maximal Galois extension of F which is unramified outside Σ.

For $G = \mathrm{GL}_{2,F}$, Hida has produced a very big Hecke algebra in [22] by the method of cohomological ℓ-adic interpolation[7]. Let $\mathcal{X}_\Sigma^{\mathrm{gl}}$ be the maximal pro-ℓ abelian quotient of G_Σ and $\mathcal{X}_\ell^{\mathrm{loc}}$ be the pro-ℓ completion of $\prod_{v|\ell} o_v^\times$. A pair $((k_\iota)_{\iota\in I_F}, w)$ consisting of an integer vector $(k_\iota)_{\iota\in I_F}$ and an integer w is called discrete type if $k_\iota \equiv w \bmod 2$ and $k_\iota \geq 2$ for all $\iota \in I_F$. A continuous character $\chi : \mathcal{X}_\Sigma^{\mathrm{gl}} \times \mathcal{X}_\ell^{\mathrm{loc}} \to \overline{\mathbb{Q}}_\ell^\times$ is algebraic of type $((k_\iota)_{\iota\in I_F}, w)$ if the quotient $\chi/\big(\chi_{\mathrm{cycle}}^{-w} \cdot \prod_{v|\ell}\prod_{\iota\in I_{F_v}} \chi_\iota^{k_\iota-2}\big)$ is of finite order. Here χ_{cycle} is the cyclotomic character, and χ_ι is the ℓ-adic character $o_v^\times \to \overline{\mathbb{Q}}_\ell^\times$ defined via the embedding ι.

Then a nearly ordinary Hecke algebra T_Σ is a finite local $o_\lambda[[\mathcal{X}_\Sigma^{\mathrm{gl}} \times \mathcal{X}_\ell^{\mathrm{loc}}]]$-algebra having the following properties:

- T_Σ is generated by the standard Hecke operators at all finite places.
- Take an ℓ-adic integer ring $o'_{\lambda'}$. For an o_λ-algebra homomorphism $f : T_\Sigma \to o'_{\lambda'}$ such that the induced o_λ-homomorphism $o_\lambda[[\mathcal{X}_\Sigma^{\mathrm{gl}} \times \mathcal{X}_\ell^{\mathrm{loc}}]] \to o'_{\lambda'}$ defines an algebraic character of discrete type $((k_\iota)_{\iota\in I_F}, w)$, there is a cuspidal $\mathrm{GL}_2(\mathbb{A}_F)$ representation π which is defined by a holomorphic Hilbert cusp form of type $((k_\iota)_{\iota\in I_F}, w)$.
- Any cuspidal representations obtained by specializations as above are nearly ordinary, that is, the action of standard Hecke operators at $v|\ell$ are ℓ-adic units.
- T_Σ is the biggest o_λ-algebra with the above properties. T_Σ is the universal ring which connects all nearly ordinary cuspidal representations.

Assume that for some π appearing from T_Σ by a specialization, the mod λ-reduction $\bar{\rho}_\pi$ of ρ_π is absolutely irreducible. Then there is a Galois representation $\rho_{T_\Sigma} : G_\Sigma \to \mathrm{GL}_2(T_\Sigma)$ which interpolates ρ_π's for various π ℓ-adically (see [56] for the ordinary Hecke algebra: the method of pseudo-representation of Wiles). The local representation $\rho_{T_\Sigma}|_{G_{F_v}}$ for $v|\ell$ is *nearly ordinary*, that is, it has an expression

$$\rho_{T_\Sigma}|_{G_{F_v}} \simeq \begin{pmatrix} \chi_{1,v} & * \\ 0 & \chi_{2,v} \end{pmatrix}$$

[7] This method was found by Shimura in [46] in the late 1960s.

for some local characters $\chi_{i,v}\colon G_{F_v} \to T_\Sigma^\times$ $(i = 1, 2)$ which are made explicit by the universal character of $\mathcal{X}_\ell^{\mathrm{loc}}$. Also $\det\rho_{T_\Sigma}|_{\mathcal{X}_\Sigma^{\mathrm{gl}}}$ is the universal character of $\mathcal{X}_\Sigma^{\mathrm{gl}}$ twisted by $\chi_{\mathrm{cycle}}^{-1}$.

Next we introduce the deformation-theoretical viewpoint due to Mazur. Let R_Σ be the universal nearly ordinary deformation ring of $\bar\rho$. We put the nearly ordinary condition at $v|\ell$ and no restrictions at other $v \in \Sigma$. There is a canonical ring homorphism $R_\Sigma \to T_\Sigma$ which is surjective (one recovers standard Hecke operators from ρ_{T_Σ}).

Theorem 3.1. *Assume the following conditions:*

1. *ℓ is odd, and $\bar\rho|_{F(\zeta_\ell)}$ is absolutely irreducible*[8].
2. *$\bar\rho|^{\mathrm{ss}}_{G_{F_v}}$ is either indecomposable, or is a sum of distinct characters for $v|\ell$ (G_{F_v}-distinguished).*

Then the nearly ordinary Hecke algebra T_Σ is a finite $o_\lambda[[\mathcal{X}_\Sigma^{\mathrm{gl}} \times \mathcal{X}_\ell^{\mathrm{loc}}]]$-flat algebra of complete intersection and is isomorphic to the universal nearly ordinary deformation ring R_Σ.

When $F = \mathbb{Q}$, this is a fundamental result of Wiles [57] supplemented by his collaboration with Taylor [55][9]. For general F, this result is due to the author and is a special case of [18]. For a precise construction of T_Σ at a finite level, see [18]. The above version of the theorem follows easily from it.

So Hida's theory of nearly ordinary Hecke algebras for $\mathrm{GL}_{2,F}$ is almost completed when the residual representation is absolutely irreducible.

We discuss some applications of the theorem. The first application is to the modularity of Galois representations[10]. The following theorem, which is a partial contribution to the modularity of elliptic curves over F (the Taniyama–Shimura conjecture) is an example:

Theorem 3.2. *Let E be an elliptic curve over F, and let $\rho_{E,\ell}$ be the associated Galois representation on the Tate module $T_\ell(E)$. Assume the following conditions:*

1. *3 splits completely in F.*
2. *$\rho_{E,3} \mod 3$ remains absolutely irreducible over $F(\zeta_3)$.*
3. *E is semi-stable at all $v|3$ and ordinary if it admits a good reduction.*

Then there is a cuspidal representation π_E of infinity type $((2, \dots, 2), -2)$ such that $\rho_{E,\ell}$ is isomorphic to ρ_{π_E} over $\overline{\mathbb{Q}}_\ell$ for any ℓ. In particular $L(E, s) = L(s, \pi_E, \mathrm{st})$ holds for the Hasse–Weil L-function $L(E, s)$ of E.

[8] There is an exceptional case when $\ell = 5$ and $[F(\zeta_5) : F] = 2$ which is not treated in [18]. This case will be discussed on another occasion.

[9] Also with a supplement by Diamond [13] to treat some exceptional local representations.

[10] One may apply the solvable base change technique to know the modularity, so only a weaker form of Theorem 3.1 is needed. Usually $[F : \mathbb{Q}]$ becomes even after a base change, so the Mazur principle in the even degree case [17] is indispensable for this approach.

One can use other division points (5-division points, for example) to establish similar modularity results.

By the solvable base change technique, one can also deduce a quite interesting result for $\mathrm{GL}_{2,\mathbb{Q}}$. See the work of Khare and Wintenberger [30], [28] for such directions. In particular Khare proves Serre's conjecture for mod ℓ level 1 forms of $\mathbb{Q}$.

The second application is in Iwasawa theory, especially the Selmer group for the symmetric square of GL_2-representations [23]. Here we need the full form of Theorem 3.1. This theorem is also effective in solving classical problems. See [24] for such an example (Eichler's integral basis problem).

To establish Theorem 3.1, what we shall do is something like proving Weber's theorem (all abelian extensions of $\mathbb{Q}$ are cyclotomic). In our case we already have some tower of Galois extensions which is explicitly described by automorphic forms via the reciprocity law, and we would like to know that it exhausts all extensions which satisfy reasonable conditions.

Let us list the main ingredients in the proof of Theorem 3.1, which are now standard. It consists of 4 steps:

- 1st step. *Level and weight optimization.* Define some minimal Hecke algebra $T_{\min}$. It is necessary to find a minimal Σ with more restrictive minimality conditions.
- 2nd step. *Compatibility of local and global Langlands correspondences.* Define the Galois representation $\rho_{T_{\min}}\colon G_F \to \mathrm{GL}_2(T_{\min})$, and determine the local behaviour of $\rho_{T_{\min}}$, namely local restrictions $\rho_{T_{\min}}|_{G_{F_v}}$.
- 3rd step. $R = T$ *in the minimal case.* Define the minimal deformation ring $R_{\min}$ and show that $R_{\min} \simeq T_{\min}$.
- 4th step. *Reduction to the minimal case.* Show $R_\Sigma \simeq T_\Sigma$ by reducing it to the minimal case (raising the level).

The basic strategy for the third and fourth steps exists for general modular varieties $M(G, X)$, and will be discussed in the next two sections.

The first step applied to $\mathrm{GL}_{2,\mathbb{Q}}$ is Serre's ε-conjecture on the optimization of level and weights for modular $\mathrm{GL}_2(\overline{\mathbb{F}}_\ell)$-representations. This is proved in a series of works. See [40] for the related references. For general $\mathrm{GL}_{2,F}$, outside ℓ, the optimization of the level is completely understood; see [26], [27], [17] for the ramified case and the Mazur principle, [36] for the Ribet type theorem [39]. These results are enough to prove Theorem 3.1 in the nearly ordinary case. However, optimization of level and weight at $v|\ell$ is still insufficient for further progress[11]. The solvable base change technique in [50] is useful for modularity questions.

For the second step, one first constructs ρ_{T_Σ} by an ℓ-adic interpolation from various ρ_π's which appear from Shimura curves by Wiles' method of pseudo-representations [56].

[11] A very naive version of Serre's ε-conjecture is false for $F \neq \mathbb{Q}$ for trivial reasons (one can not attain a conductor which is prime to ℓ in general). Diamond has formulated a refined conjecture and is making progress.

For $\mathrm{GL}_{2,F}$, $\rho_\pi|_{G_{F_v}}$ for $v \nmid \ell$ depends only on the v-component π_v of $\pi = \otimes_w \pi_w$ and is obtained by the local Langlands correspondence for GL_{2,F_v} [5], [52]. For $v|\ell$, the Weil–Deligne representation attached to the potentially stable representation $\rho_\pi|_{G_{F_v}}$ is compatible with π_v when ρ_π comes from Shimura curves [41], [42]. However, this information is too weak to determine $\rho_\pi|_{G_{F_v}}$. For example, information on the Hodge filtration is lacking. Breuil has formulated a p-adic version of the local Langlands correspondence [2].

For general Shimura varieties, this requires a detailed study of bad reductions. For n-dimensional representations as defined by Clozel [9], compatibility outside ℓ is established. (See [20] for the semi-simplification case, and see Taylor and Yoshida [54] for the general case).

For $v|\ell$, p-adic Hodge theory is the basic tool at present, especially for general Shimura varieties.

Remark 3.3. 1. One may allow finite flat deformations at $v|\ell$ when F_v is absolutely unramified [18]. Kisin has developed a local theory for finite flat deformation which allows ramification when the residue field at v is $\mathbb{F}_\ell$, with an application to modularity theorems [31].

2. In the case of $\mathbb{Q}$, Ramakrishna and Khare [29] have found an interesting method without any reduction to the minimal case via use of special deformations, that is, by adding Steinberg type conditions at several auxiliary primes[12]. Moreover, Ramakrishna has studied various Galois deformations over $\mathbb{Q}$, which may not be modular a priori, .

When $\bar\rho$ is absolutely reducible, T_Σ is used to prove Iwasawa's main conjecture for class characters of totally real fields. Skinner and Wiles [49], [51] have obtained a modularity result even in the absolutely reducible case, by showing partial results on the relationship between the versal hull of R_Σ and T_Σ using many ideas and techniques including Taylor–Wiles systems.

4. Taylor–Wiles systems: the formalism

In [55], Taylor and Wiles have invented a marvelous argument to show a nice ring theoretical property of $T_{\min}$: $T_{\min}$ is a complete intersection in the case of modular curves. This property may seem technical at first glance, but has the very deep implication that $R_{\min}$ is $T_{\min}$ in the case they considered. The original argument by Taylor and Wiles has been improved by now; Faltings suggested the direct use of deformation rings rather than Hecke algebras, and the further refinement we now describe is due to Diamond [14] and myself.

For a number field F, let $|F|_f$ be the set of finite places of F. q_v denotes the cardinality of the residue field $k(v)$ for $v \in |F|_f$.

[12] A freeness assertion is necessary for this approach. Using this assertion, one can compute congruence modules to come back to unrestricted deformations. Technical assumptions are made on $\bar\rho$ in [29].

Let k_λ be the residue field of o_λ.

Definition 4.1. Let H be a torus over F of relative dimension d, and let X be a set of finite subsets of $|F|_f$ that contains $\emptyset$. Let R be a complete noetherian local o_λ-algebra with the residue field k_λ, and M an R-module which is finitely generated as an o_λ-module. A Taylor–Wiles system $\{R_Q, M_Q\}_{Q\in X}$ for (R, M) consists of the following data:

- For $Q \in X$ and $v \in Q$, H is split at v, and $q_v \equiv 1 \mod \ell$. We denote the ℓ-Sylow subgroup of $H(k(v))$ by Δ_v, and Δ_Q is defined as $\prod_{v\in Q} \Delta_v$ for $Q \in X$.
- For $Q \in X$, R_Q is a complete noetherian local $o_\lambda[\Delta_Q]$-algebra with the residue field k_λ, and M_Q is an R_Q-module. For $Q = \emptyset$, $(R_\emptyset, M_\emptyset) = (R, M)$.
- There is a surjection of local o_λ-algebras

$$R_Q/I_Q R_Q \to R$$

  for each non-empty $Q \in X$ [13]. Here $I_Q \subset o_\lambda[\Delta_Q]$ denotes the augmentation ideal of $o_\lambda[\Delta_Q]$.
- The homomorphism $R_Q/I_Q R_Q \to \mathrm{End}_{o_\lambda} M_Q/I_Q M_Q$ factors through R, and $M_Q/I_Q M_Q$ is isomorphic to M as an R-module.
- M_Q is free and of rank α as an $o_\lambda[\Delta_Q]$-module for a fixed $\alpha \geq 1$.

In [55], the condition that R_Q is Gorenstein and M_Q is a free R_Q-module is required.

Taylor–Wiles systems are commutative ring theoretic versions of Kolyvagin's Euler systems. The following theorem is the most important consequence of the existence of Taylor–Wiles systems.

Theorem 4.2 (Complete intersection and freeness criterion). *For a Taylor–Wiles system $\{R_Q,\ M_Q\}_{Q\in X}$ for (R, M) and a torus H of dimension d, assume the following conditions:*

1. *For any $m \in \mathbb{N}$*
$$v \in Q \Rightarrow q_v \equiv 1 \mod \ell^m$$
holds for infinitely many $Q \in X$.
2. *The cardinality r of Q is independent of $Q \in X$ for non-empty Q.*
3. *R_Q is generated by at most dr elements as a complete local o_λ-algebra for non-empty $Q \in X$.*

[13] For deformation rings, $R_Q/I_Q R_Q \simeq R$ usually holds. The condition here is relaxed so that it applies to Hecke algebras directly.

Then one has:

- (*complete intersection property*) *R is o_λ-flat and of relative complete intersection of dimension zero;*
- (*freeness*) *M is a free R-module.*

In particular, M is a faithful R-module. If we denote the image of R in $\mathrm{End}_{o_\lambda} M$ by T, we have $R \simeq T$ as a consequence of Theorem 4.2. So the complete intersection and the freeness criterion is also regarded as an *isomorphism criterion*. The Auslander–Buchsbaum formula for regular local rings plays a very important role in the proof of the theorem.

The reason why we have the complete intersection property is the following. When Q and N vary, a well chosen limit of $R_Q/m_{R_Q}^N$ tends to be a power series ring over o_λ in dr variables[14]. This is implied by condition (3) of 4.2. Usually $R = R_Q/I_Q R_Q$ holds, thus R is defined by $d \cdot \sharp Q = dr$ equations in R_Q, and hence it is a complete intersection in the limit.

If the complete intersection property and freeness both hold, then this pair of properties is inherited by descendants of (R, M). This is formulated in the following way.

Definition 4.3. 1. An admissible quintet is a quintet $(R, T, \pi, M, \langle\ ,\ \rangle)$, where R is a complete local o_λ-algebra, T is a finite flat o_λ-algebra, $\pi : R \to T$ is a surjective o_λ-algebra homomorphism, M is a faithful finitely generated T-module which is o_λ-free, and $\langle\ ,\ \rangle : M \otimes_{o_\lambda} M \to o_\lambda$ is a perfect pairing which induces $M \simeq \mathrm{Hom}_{o_\lambda}(M, o_\lambda)$ as a T-module.

2. An admissible quintet $(R, T, \pi, M, \langle\ ,\ \rangle)$ is distinguished if R is a complete intersection and M is R-free (and hence π is an isomorphism).

3. By an admissible morphism from $(R', T', \pi', M', \langle\ ,\ \rangle')$ to $(R, T, \pi, M, \langle\ ,\ \rangle)$ we mean a triple (α, β, ξ). Here $\alpha : R' \to R$, $\beta : T' \to T$ are surjective o_λ-algebra homomorphisms making the following diagram

$$\begin{array}{ccc} R' & \xrightarrow{\alpha} & R \\ {\scriptstyle \pi'}\downarrow & & \downarrow{\scriptstyle \pi} \\ T' & \xrightarrow{\beta} & T \end{array}$$

commutative, and $\xi : M \hookrightarrow M'$ is an injective T'-homomorphism onto an o_λ-direct summand. (Note that we do not assume the restriction of $\langle\ ,\ \rangle'$ to $\xi(M)$ is $\langle\ ,\ \rangle$.)

A Taylor–Wiles system gives rise to a distinguished admissible quintet for a suitably chosen pairing on M under the conditions of Theorem 4.2.

[14] In applications, R_Q is a deformation ring. This implies that the deformation functor which defines R_Q behaves as if it were unobstructed for a well-chosen limit of Q.

Assume that (α, β, ξ) is an admissible morphism from $(R', T', \pi', M', \langle\ ,\ \rangle')$ to $(R, T, \pi, M, \langle\ ,\ \rangle)$. There is an abstract criterion for $(R', T', \pi', M', \langle\ ,\ \rangle')$ to be a distinguished quintet if $(R, T, \pi, M, \langle\ ,\ \rangle)$ is distinguished.

By duality we have

$$\hat{\xi}\colon M' \simeq \mathrm{Hom}_{o_\lambda}(M', o_\lambda) \to \mathrm{Hom}_{o_\lambda}(M, o_\lambda) \simeq M$$

such that

$$\langle\, \xi(x),\ y \,\rangle' = \langle\, x,\ \hat{\xi}(y) \,\rangle \quad \text{for all } x \in M,\ y \in M'.$$

We fix an o_λ-algebra homomorphism $f_T\colon T \to o_\lambda$. We define f_R (resp. $f'_{R'}$) as $f_T \circ \pi$ (resp. $f \circ \beta \circ \pi'$).

Theorem 4.4 (abstract level raising formalism). *For an admissible morphism between admissible quintets $(R', T', \pi', M', \langle\ ,\ \rangle') \to (R, T, \pi, M, \langle\ ,\ \rangle)$, we assume the following conditions:*

1. *$(R, T, \pi, M, \langle\ ,\ \rangle)$ is distinguished.*
2. *T and T' are reduced, $M' \otimes_{o_\lambda} E_\lambda$ is $T' \otimes_{o_\lambda} E_\lambda$-free, and its rank is the same as the rank of M over T.*
3. *$\hat{\xi} \circ \xi(M) = \Delta \cdot M$ holds for some non-zero divisor Δ in T.*
4. *An inequality*

$$\begin{aligned}&\mathrm{length}_{o_\lambda} \ker f_{R'}/(\ker f_{R'})^2\\ &\quad \le \mathrm{length}_{o_\lambda} \ker f_R/(\ker f_R)^2 + \mathrm{length}_{o_\lambda} o_\lambda/\ f_T(\Delta)o_\lambda\end{aligned}$$

holds.

Then $(R', T', \pi', M', \langle\ ,\ \rangle')$ is also distinguished, that is, $R' \simeq T'$, R is a complete intersection, and M' is T'-free.

A generalization of Wiles' isomorphism criterion in [57] by Lenstra is used in the proof. Though the main two theorems in this section are a consequence of the general commutative algebra machinery, they are extremely deep. In the next section, we will see how modular varieties should yield the setup formulated in this section.

Remark 4.5. 1. The idea of an admissible quintet first appeared in the work of Ribet in [38]. $M/\hat{\xi} \circ \xi(M)$ is the congruence module.

2. The o_λ-direct summand property in Definition 4.3, (3) is an abstract form of what is known as "Ihara's Lemma".

3. $\ker f_R/(\ker f_R)^2$ is regarded as a Selmer group. When R is a complete intersection, its o_λ-length is computed and equal to $\mathrm{length}_{o_\lambda} o_\lambda/\eta_R$. Here η_R is the ideal generated by the image of 1 under $o_\lambda \xrightarrow{\hat{f}_R} \mathrm{Hom}_{o_\lambda}(R, o_\lambda) \simeq R \xrightarrow{f_R} o_\lambda$ as introduced by Wiles in [57]. Note that the finiteness of the Selmer group is a consequence of the reducedness of T (Taylor–Wiles systems do not give finiteness directly).

5. Taylor–Wiles systems: a strategy for the construction

Take a pair (G, X) which defines a modular variety, and take a finite dimensional representation $\nu\colon G \to \mathrm{Aut}_{E_\lambda} V_{E_\lambda}$ defined over E_λ. For simplicity, we assume that G is quasi-split over $\mathbb{Q}$. ν defines a $G(\mathbb{A}_f)$-equivariant local system $\mathcal{F}_\nu$ on $M(G, X)$ [15], which yields a family of local systems $\mathcal{F}_\nu^K$ on $M_K(G, X)$ at each level K. By choosing an o_λ-lattice in V_{E_λ}, $\mathcal{F}_\nu^K$ has an o_λ-structure $\mathcal{F}_{\nu,o_\lambda}^K$. We limit ourselves to the adelic action and the Hecke algebra action which respect the integral structure. Let Σ be a finite set of primes containing ℓ, and for a factorizable $K = \prod_q K_q$, K_Σ and K^Σ are defined by $K_\Sigma = \prod_{q\in\Sigma} K_q$, $K^\Sigma = \prod_{q\notin\Sigma} K_q$.

For simplicity, we only consider Hecke operators outside Σ [16]. $H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda}$ is the Hecke algebra formed by o_λ-valued functions on the adele group without components in Σ. Let us begin with a remark on homological algebras: $R\Gamma_B(M_K(G, X), \mathcal{F}_{\nu,o_\lambda}^K)$ belongs to $D^+(H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda})$, the derived category of $H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda}$-complexes bounded from below. This is shown by taking the canonical flasque resolution of $\mathcal{F}_{\nu,o_\lambda}^K$.

Assume that K_q is maximal and hyperspecial for $q \notin \Sigma$. Then $H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda} = \otimes_{q\notin\Sigma} H(G(\mathbb{Q}_q), K_q)_{o_\lambda}$ is commutative. Let m_Σ denote a maximal ideal of $H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda}$. Throughout this section we make the following assumption:

Assumption 5.1 (Vanishing of cohomology for $\overline{\mathbb{F}}_\ell$-coefficients). For any (sufficiently small) K,

$$H_B^i(M(G, X)_K, \mathcal{F}_{\nu,o_\lambda}^K \otimes_{o_\lambda} k_\lambda)_{m_\Sigma} = 0$$

for $i \neq q(G)$.

This type of vanishing statement is known over E_λ when the weight of ν is sufficiently regular, but for torsion coefficients it is difficult to prove a vanishing statement[17]. Moreover, we need to have a vanishing claim *which holds uniformly* in K.

Let $M_\Sigma = H_B^{q(G)}(M(G, X)_K, \mathcal{F}_{\nu,o_\lambda}^K)_{m_\Sigma}$ be the localization of the middle dimensional cohomology group at m_Σ[18]. By Assumption 5.1, it is o_λ-free.

T_Σ is defined as the image of $H(G(\mathbb{A}_f^\Sigma), K^\Sigma)_{o_\lambda}$ in M_Σ. We call it the ℓ-adic Hecke algebra.

When $M(G, X)$ is a Shimura variety $\mathrm{Sh}(G, X)$ with reflex field E, we further assume, by using the étale cohomology, that a reciprocity law of the following form for $(\mathrm{Sh}(G, X), \nu)$ holds:

$$M_\Sigma \otimes_{o_\lambda} \overline{\mathbb{Q}}_\ell \simeq \oplus_{\pi\in\Pi_\rho} V_\rho \otimes (\pi_f^K)^{a(\pi)}.$$

[15] For the center $Z(G)$ of G, the image of $Z(G)(\mathbb{Q}) \cap K$ by ν must be trivial for some K.

[16] We also need Hecke actions at Σ. At ℓ, one should use a semi-subgroup $G(\mathbb{Q}_\ell)_+$ of $G(\mathbb{Q}_\ell)$ because the full action of $G(\mathbb{Q}_\ell)$ does not preserve the lattice structure in general.

[17] In fact, there are non-zero torsion classes in general unless one puts restrictive conditions on m_Σ.

[18] Further localizations by elements in $H(G(\mathbb{A}_\Sigma), K_\Sigma)$ are necessary.

Here ${}^L G$ is the L-group of G, Π_ρ is a packet which corresponds to $\rho\colon G_E \to {}^L G(\overline{\mathbb{Q}}_\ell)$ and is cohomological for ν [19]. V_ρ is a G_E-representation defined by a finite dimensional representation of ${}^L G$ determined by ν. We try to construct a Taylor–Wiles system for (T_Σ, M_Σ).

As is suggested in the introduction, a modular variety for a given level K does not admit a symmetry by a group. To have a group action, we must use an open subgroup of $G(\mathbb{A}_f)$ which is smaller than K. Take a $\mathbb{Q}$-parabolic subgroup P of G, and a quotient torus H of P which is disjoint from the center $Z(G)$, that is, the image of $Z(G)$ in H is trivial. It is very important to have a torus which is disjoint from the center: this gives us a *geometric direction*, which, in the case of Shimura varieties, cannot be obtained from abelian extensions of the reflex field. This feature is quite different from Euler systems.

For a prime q, let $K_{P,q} = \{g \in G(\mathbb{Z}_q),\ g \bmod q \in P(\mathbb{F}_q)\}$ be a parahoric subgroup at q defined by P, and let $K_{P,H,q} = \{g \in K_{P,q},\ g \bmod q \in \ker(P(\mathbb{F}_q) \to H(\mathbb{F}_q)^\ell)\}$ be a subgroup depending on H. Here $H(\mathbb{F}_q)^\ell$ is the maximal subgroup whose order is prime to ℓ. Then $K_{P,q}/K_{P,H,q}$ is isomorphic to Δ_q, the ℓ-Sylow subgroup of $H(\mathbb{F}_q)$.

For a finite set of finite primes Q which is disjoint from Σ, one sets

$$K_{P,Q} = \prod_{q \in Q} K_{P,q} \cdot K^Q,$$

$$K_{P,H,Q} = \prod_{q \in Q} K_{P,H,q} \cdot K^Q,$$

and $\Delta_Q = \prod_{q \in Q} \Delta_q$. Since H is disjoint from the center, the natural covering

$$\pi_Q\colon \mathrm{Sh}_{K_{P,H,Q}} \to \mathrm{Sh}_{K_{P,Q}}$$

is an étale Galois covering with Galois group Δ_Q if K^Q is small enough to make group actions free. We view $R\Gamma_B(\mathrm{Sh}_{K_{P,H,Q}} \mathcal{F}_{\nu,o_\lambda}^{K_{P,H,Q}})$ as an object in

$$D^+(H(G(\mathbb{A}_f^{\Sigma \cup Q}), K^{\Sigma \cup Q})_{o_\lambda}).$$

Then we make the following simple observation (*perfect complex argument*):

Lemma 5.2. *Let $\pi\colon X \to Y$ be an étale Galois covering between finite dimensional manifolds*[20] *with Galois group G. Let $\mathcal{F}$ be a smooth Λ-sheaf on Y. Then $R\Gamma_B(X, \pi^*\mathcal{F})$ is represented by a perfect complex of $\Lambda[G]$-modules, and*

$$R\Gamma_B(X, \pi^*\mathcal{F}) \otimes^{\mathbb{L}}_{\Lambda[G]} \Lambda[G]/I_G \simeq R\Gamma_B(Y, \mathcal{F})$$

holds in $D^b(\Lambda[G])$. Here I_G is the augmentation ideal of $\Lambda[G]$, and the map is induced by the trace map.

[19] We are ignoring the role of characters of centralizer groups, endoscopy, and so on. See [32] for a precise conjectural description. Moreover, we only have information on the semi-simplification V_ρ^{ss}.

[20] These manifolds must satisfy reasonable finiteness conditions on cohomology groups, which are true for the algebraic varieties or modular varieties that we are interested in.

By Lemma 5.2, we know that $R\Gamma_B(\mathrm{Sh}_{K_{P,H,Q}}, \mathcal{F}_{\nu,o_\lambda}^{K_{P,H,Q}})$ is, in $D^b(o_\lambda[\Delta_Q])$, a perfect complex of $o_\lambda[\Delta_Q]$-modules, that is, it is quasi-isomorphic to a bounded complex of free $o_\lambda[\Delta_Q]$-modules. We assume 5.1 holds and localize at the maximal ideal m_Q of $H(G(\mathbb{A}_f^{\Sigma\cup Q}), K^{\Sigma\cup Q})_{o_\lambda}$ below m_Σ. Then

$$M_Q = H_B^{q(G)}(\mathrm{Sh}_{K_{P,H,Q}}, \mathcal{F}_{\nu,o_\lambda}^{K_{P,H,Q}})_{m_Q}$$

is $o_\lambda[\Delta_Q]$-free since it is the only non-zero cohomology of a perfect $o_\lambda[\Delta_Q]$-complex. Then $M_Q \otimes_{o_\lambda[\Delta_Q]} o_\lambda[\Delta_Q]/I_{\Delta_Q} = M_{0,Q}$. Here $M_{0,Q} = H_B^{q(G)}(\mathrm{Sh}_{K_{P,Q}}, \mathcal{F}_{\nu,o_\lambda}^{K_{P,Q}})_{m_Q}$.

We define T_Q as the image of $H(G(\mathbb{A}_f^{\Sigma\cup Q}), K^{\Sigma\cup Q})_{o_\lambda}$ in $\mathrm{End}_{o_\lambda} M_Q$. Then the expectation is that $M_{0,Q}$ is a direct sum of M for a good choice of Q, and that (T_Q, M_Q) forms a Taylor–Wiles system for (T_Σ, M_Σ) as Q varies. Hence, a system for deformation rings will be obtained for appropriate choices of deformation functors.

To relate $M_{0,Q}$ to the original M_Σ, we already need information from the Galois parameter for the Langlands (or Arthur) packets for automorphic forms which contributes to $M_{0,Q}$, since we need to remove non-spherical components from $M_{0,Q}$. Some version of the compatibility of local and global parametrizations is needed.

This program for constructing a Taylor–Wiles system, and in particular for combining the perfect complex argument and the vanishing assumption 5.1 to obtain a system, was made explicit and carried out in two special cases in [18]. One case is when (G, X) defines a Shimura curve, that is, $G = G_D$ as in §3. The Hecke actions on H_B^0 and H_B^2 are easily determined for Shimura curves, and Assumption 5.1 is true if we localize at the maximal ideals of Hecke algebras which correspond to absolutely irreducible two dimensional representations. In another case, Hida varieties are used to treat some situations which do not arise from Shimura curves.

In general, the program is very effective when G defines Hida varieties, since the vanishing assumption is trivially true. To study non-abelian class field theory, especially the question of modularity, or deformations of absolutely irreducible representations, this case is quite useful, since one can apply the Jacquet–Langlands correspondence for inner forms (assuming that it is already known) to convert the problem to a compact inner form. For unitary groups, this approach was taken by Harris and Taylor, showing the $R = T$ theorem for the n-dimensional representations of CM fields constructed by Clozel [9] under several restrictive assumptions. Arithmetic geometrical difficulties are all encoded in the compatibility of local and global Langlands correspondences outside ℓ and in the description of local monodromy at ℓ.

Unfortunately, Hida varieties do not admit any natural Galois action. Since one needs to have finer information coming from geometry, it is still an important problem to construct Taylor–Wiles systems for higher dimensional Shimura varieties.

There are several possible solutions. One solution would be to establish Assumption 5.1, that is, a vanishing theorem for $\overline{\mathbb{F}}_\ell$-coefficients. A partial solution will be given in the sequel for unitary groups with signature $(m, 1)$ by using a geometrical realization of the Jacquet–Langlands correspondence. As far as the author knows,

these unitary Shimura varieties are the only examples where the program is carried out for the middle dimensional cohomology in arbitrary higher dimensions.

For other approaches in the case of Siegel modular varieties using p-adic Hodge theory, see [34]. A Taylor–Wiles system for the Hilbert–Siegel case, especially $G\,\mathrm{Sp}_{4,\mathbb{Q}}$, has been studied by Tilouine.

Another possible solution would be to avoid the use of vanishing theorems. We have enough information about the alternating sum of cohomology groups, so in working with a suitable K_0-group of virtual Galois–Hecke bimodules instead of Galois–Hecke bimodules, Taylor–Wiles systems might work for virtual representations.

Remark 5.3. The perfect complex argument is simple but very powerful in the cohomological study of congruences between automorphic forms. There are several other uses of the perfect complex argument.

One example would be the construction of nearly ordinary Hecke algebra for GL_2, with an *exact control theorem*.

In particular for the nearly ordinary Hecke algebra T_Σ, there is a faithful T_Σ-module M_Σ with the following properties:

- M_Σ is free over $\Lambda = o_\lambda[[\mathcal{X}_\Sigma^{\mathrm{gl}} \times \mathcal{X}_\ell^{\mathrm{loc}}]]$.
- Take an algebraic character $\chi : \Lambda \to o'_{\lambda'}$ of discrete type $((k_\iota)_{\iota\in I_F}, w)$ such that $\tilde{\chi} = \chi/(\chi_{\mathrm{cycle}}^{-w} \cdot \prod_{v|\ell} \prod_{\iota\in I_{F_v}} \chi_\iota^{k_\iota-2})$ is of order prime to ℓ. $M_\Sigma \otimes_{\Lambda,\chi} o'_{\lambda'}$ is a lattice in the space of nearly ordinary forms of type $((k_\iota)_{\iota\in I_F}, w)$ with "nebencharacter" $\tilde{\chi}$. Moreover, when the degree $[F : \mathbb{Q}]$ is even[21], it is *exactly* the image of $H_B^0(M_K, \mathcal{F}^K_{((k_\iota)_{\iota\in I_F},w),o'_{\lambda'}})$. Here M_K is a Hida variety associated to a definite quaternion algebra, and $\mathcal{F}^K_{((k_\iota)_{\iota\in I_F},w),o'_{\lambda'}}$ is a smooth $o'_{\lambda'}$-sheaf on M_K (cf. [17]) which realizes the reciprocity law for forms of type $((k_\iota)_{\iota\in I_F}, w)$ over $\overline{\mathbb{Q}}_\ell$.

Another example would be the level optimization for GL_2 in a special case, which gives an interpretation of Carayol's lemma [7], [17].

6. Geometric Jacquet–Langlands correspondence

We use the strategy in Section 5 for some unitary Shimura varieties where the non-abelian reciprocity is established by Kottwitz [33], and we construct a Taylor–Wiles system for the middle dimensional cohomology group. To do this, we establish an explicit Jacquet–Langlands correspondence which preserves o_λ-lattice structures by arithmetic geometrical means, in particular by analyzing bad reductions of Shimura varieties. This is called a geometric Jacquet–Langlands correspondence. This first

[21] One uses Shimura curves when the degree is odd.

appeared in the level optimization problem mentioned in §3 in the case of modular curves. This also gives an arithmetic geometrical meaning of Hida varieties.

Let us give a simple example. Let $S_{D,K} = \mathrm{Sh}(G_D, X_D)_K$ be a Shimura curve as in §2. For a finite place v of F, assume that D is split at v, and $v \nmid \ell$. Let $\tilde{D}$ be a quaternion algebra which is definite at all infinite places, non-split at v, but has the same local invariants as D at the other finite places. $\tilde{D}$ defines a Hida variety $M_{\tilde{D}}$, and we have an isomorphism $D^\times(\mathbb{A}^v_{F,f}) \simeq \tilde{D}^\times(\mathbb{A}^v_{F,f})$ outside v.

Assume $K = \prod_u K_u$ is factorizable, and K_v is an Iwahori subgroup of GL_{2,F_v}. We set $\tilde{K}_v = o^\times_{\tilde{D}_v}$ for a maximal order $o_{\tilde{D}_v}$ of D_v, and we view $\tilde{K} = \tilde{K}_v \cdot \prod_{u\neq v} K_v$ as a subgroup of $\tilde{D}^\times(\mathbb{A}_{F,f})$.

Then we have the following geometric realization of the Jacquet–Langlands [25], and of the Shimizu [43] correspondence[22]:

Proposition 6.1. *For $\Lambda = E_\lambda$, we have a (non-canonical) isomorphism as $H(D^\times(\mathbb{A}^{\ell,v}_{F,f}), K^{\ell,v})_{o_\lambda}$-modules*

$$W_0 H^1_{\text{ét}}(S_{D,K}, \Lambda) \simeq H^0(M_{\tilde{D},\tilde{K}}, \Lambda)$$

modulo Eisenstein modules. W_0 is the weight 0 part of the weight filtration, and the decomposition with respect to the Hecke action gives the Jacquet–Langlands correspondence.

Here we view the Hida variety $M_{\tilde{D},\tilde{K}}$ as a variety over $\overline{k(v)}$, and we regard it as the set of supersingular points[23], that is, the set of points which correspond to formal o_v-modules of height 2. Since the special fiber at v of the arithmetic model of $S_{D,K}$ has two kinds of components which meet at supersingular points, the claim follows from the standard calculation of the weight filtration using the dual graph of the special fiber[24]. There is also a variant for any finite o_λ-algebra Λ, which preserves o_λ-lattice structures.

So the Jacquet–Langlands correspondence is realized on the weight filtration of local monodromy action where the Shimura variety admits a bad reduction. This is our starting point[25].

6.1. Arithmetic model of unitary Shimura varieties. As in §2, F denotes a totally real field and $[F : \mathbb{Q}] = g$. I_F is the set of the embeddings $\iota\colon F \hookrightarrow \mathbb{R}$. Take an imaginary quadratic field E_0 over $\mathbb{Q}$, and let $E = E_0 \cdot F$ be the composite field. Let $\mathrm{Gal}(E/F) = \langle\sigma\rangle$.

Let D be a central division algebra over E of dimension n^2, and let $*\colon D \to D$ be a positive involution of the second kind, that is, an involution which induces σ on E.

[22] There is also a version when D is ramified at v, which plays an essential role in [39], [36].

[23] This identification is non-canonical.

[24] The Hecke action on the set of the irreducible components is Eisenstein, and we are ignoring this part.

[25] It will be desirable to have a motivic correspondence over a global field.

We require the following condition on ι at infinity: At one infinite place $\iota_0 \in I_F$, $*$ is equivalent to $g \mapsto J^t \bar{g} J^{-1}$ with $J = \begin{pmatrix} 1_{n-1} & 0 \\ 0 & -1 \end{pmatrix}$, and it is equivalent to the standard involution $g \mapsto^t \bar{g}$ for the other $\iota \neq \iota_0$.

Let $U(D) = \{g \in D^{\mathrm{op}\times},\ g \cdot g^* = 1_D\}$ be the unitary group. Let $GU(D) = \{g \in D^{\mathrm{op}\times},\ g \cdot g^* = \nu(g) \cdot 1_D,\ \nu(g) \in \mathbb{G}_{m,F}\}$ be the group of unitary similitudes, seen as a reductive group over F.

Let $G' = \mathrm{Res}_{F/\mathbb{Q}} U(D)$ be the Weil restriction of the unitary group, and let G be the inverse image of $\mathbb{G}_{m,\mathbb{Q}}$ by $\mathrm{Res}_{F/\mathbb{Q}} GU(D) \to \mathrm{Res}_{F/\mathbb{Q}} \mathbb{G}_{m,F}$.

$$1 \to G' \to G \xrightarrow{\nu} \mathbb{G}_{m,\mathbb{Q}} \to 1.$$

We fix an embedding $E_0 \hookrightarrow \mathbb{C}$. This defines a CM-type on E, and for each $\iota\colon F \hookrightarrow \mathbb{R}$, we have identifications

$$E \otimes_{F,\iota} \mathbb{R} \simeq \mathbb{C},\ D \otimes_{F,\iota} \mathbb{R} \simeq M_n(\mathbb{C}).$$

It follows that

$$G'_{\mathbb{R}} \simeq U(n-1,1) \times U(n)^{g-1}$$

by our choice.

Define a group homomorphism

$$h_0\colon \mathrm{Res}_{\mathbb{C}/\mathbb{R}} \mathbb{G}_{m,\mathbb{C}} \to G_{\mathbb{R}}$$
$$z \mapsto h_0(z) = \left(\begin{pmatrix} z \cdot 1_{n-1} & 0 \\ 0 & \bar{z} \end{pmatrix}, z1_n, \ldots, z1_n \right).$$

The associated symmetric space

$$X = G_\infty(\mathbb{R})/K_\infty$$

(K_∞ is the centralizer of $h_0(\sqrt{-1})$) is a complex ball of dimension $n-1$.

For a compact open subgroup $K \subset G(\mathbb{A}_f)$, the corresponding Shimura variety $\mathrm{Sh}_K(G,X)$ is compact by our assumption on D, and the reflex field $E(G,X)$ is E. We view $\mathrm{Sh}_K(G,X)$ as a generalization of Shimura curves. The determination of the reciprocity law is due to Kottwitz [33], and the bad reductions have been studied especially by Rapoport and Zink [37], Harris, and Taylor [20].

The moduli interpretation of these varieties is given by Shimura. We need arithmetic models over the integers, so we consider it over $o_E \otimes_{\mathbb{Z}} \mathbb{Z}_p$ by fixing a prime p. The details are found in [20]. To avoid technical problems which arise from obstructions to the Hasse principle for G, we assume that n is odd in this exposition.

We fix a finite place w of the reflex field E of residual characteristic p. Assume p splits completely in E_0, and let $\wp$ be the place of E_0 below w. $P_\wp$ is the set of places of E which divide $\wp$.

Since p is split in E_0, the unitary group has a simpler form, and we have an isomorphism

$$G_{\mathbb{Q}_p} \simeq \mathbb{G}_{m,\mathbb{Q}_p} \times \prod_{u \in P_\wp} \mathrm{Res}_{E_u/\mathbb{Q}_p} D_u^{\mathrm{op}\times}.$$

In the following, we only consider compact open subgroups K_p of $G(\mathbb{Q}_p)$ which are of the form $\mathbb{Z}_p^\times \cdot \prod_{u\in P_\wp} K_u$ for $K_u \subset D^{\mathrm{op}\times}(E_u)$. We assume moreover that D is split at w, and we identify the w-component with $H_w = D_w^{\mathrm{op}\times} \simeq \mathrm{GL}_{n,E_w}$.

Take a maximal order o_D of D which is stable under $*$. $o_D \otimes_{\mathbb{Z}} \mathbb{Z}_p$ is identified with $\prod_{u\in P_\wp} o_{D_u} \times \prod_{u\in P_\wp} o_{D_{\sigma(u)}}$. For $u \in P_\wp$, let e_u denote the projector corresponding to o_{D_u}.

For an o_w-scheme U, let A be an abelian scheme over U of relative dimension gn^2 with o_D-multiplication. We assume that the o_D-action on the relative Lie algebra $\mathrm{Lie}A/U$ satisfies the following conditions:

1. For $u \in P_\wp$ different from w, $e_u\mathrm{Lie}A/U$ is zero.
2. $e_w\mathrm{Lie}A/U$ is a locally free o_U-module of rank n, and the o_w-action on $e_w\mathrm{Lie}A/S$ is the multiplication through $o_w \to \Gamma(U, \mathcal{O}_U)$.

By condition (1), for the p-divisible group $A[p^\infty]$, $e_uA[p^\infty]$ is an étale o_u-divisible module for $u \in P_\wp$ different from w. We denote the Tate module by $T_u(A)$.

We take a compact open subgroup $K = K_p \cdot K^p \subset G(\mathbb{A}_f)$. We assume that $K_w = \mathrm{GL}_n(o_w)$ and that K^p is sufficiently small. We denote the product $\mathbb{Z}_p^\times \cdot \prod_{u\in P_\wp, u\neq w} K_u \cdot K^p$ by K^w. Then $K = K_w \cdot K^w$ by the definition. We view D as a left D^{op}-module and denote it by V [26]. The integral structure is given by $V_{\mathbb{Z}} = o_D$.

Definition 6.2. Let U be an o_w-scheme, let A be an abelian scheme over U with o_D-multiplication which satisfies the Lie algebra conditions described above, and let $\bar{p}$ be a homogeneous polarization of A. Then $(A, \bar{p})$ is of type $(o_D, V_{\mathbb{Z}}; K)$ if the following rigidification structures are attached:

1. For any point $s \in U$, there is a polarization $\lambda \in \bar{p}$ of degree prime to p such that λ induces $*$ on D as its Rosati involution.
2. For any geometric point $\bar{s}$ of characteristic p, a class $k \mod K^w$ of o_D-linear symplectic similitudes

$$k: \prod_{u\in P_\wp, u\neq w} T_u(A)_{\bar{s}} \times T^p(A)_{\bar{s}} \simeq V_{\mathbb{Z}} \otimes \hat{\mathbb{Z}}^p.$$

Here $\pi_1(U, \bar{s})$, the étale fundamental group, is mapped to K^w.

By a standard argument, the moduli functor

$$U \mapsto \mathcal{S}h_K(U),$$

where $\mathcal{S}h_K(U)$ is the set of the isomorphism classes of polarized abelian schemes of type $(o_D, V_{\mathbb{Z}}; K)$ over U, is representable by a projective smooth scheme $\mathcal{S}h_K$ over o_w, which gives a model for $\mathrm{Sh}_K(G, X)$.

[26] With a symplectic form which we do not make precise here.

Let $\mathcal{A}^{\mathrm{univ}}$ be the universal abelian scheme over $\mathcal{S}h_K$. Then o_{D_w} acts on $e_w\mathcal{A}^{\mathrm{univ}}[p^\infty]$, and is isomorphic to $M_n(o_w)$. By Morita equivalence, this gives an o_w-divisible module $\mathcal{C}_w$ of height n on $\mathcal{S}h_K$, so that $e_w\mathcal{A}^{\mathrm{univ}}[p^\infty] \simeq \mathcal{C}_w^{\oplus n}$.

For general $K_w \subset H_w(E_w) \simeq \mathrm{GL}_n(E_w)$, we have a flat arithmetic model $\mathcal{S}h_K$ over o_w by using the notion of level structures due to Drinfeld [15].

6.2. Inductive structure. We briefly describe an inductive structure of arithmetic models in characteristic p. The inductive structure was first studied by Boyer in the function field case [1]. In our unitary case it is due to Harris and Taylor in [20].

For $N \geq 1$, set $K_N = K_{N,w} \cdot K^w$, $K_{N,w} = \ker(H_w(o_w) \to H_w(o_w/m_w^N))$, $X_N = \mathcal{S}h_{K_N}$, and $S = \operatorname{Spec} o_w$. Let s (resp. η) denote the closed (resp. generic) point of S.

For $1 \leq h \leq n$, $X_{N,s}^{[h]}$ is the set of points in the special fiber $(X_N)_s$ of X_N where the étale (resp. connected) part of $\mathcal{C}_w$ has height $n-h$ (resp. height h). $X_{N,s}^{[h]}$ is locally closed in $(X_N)_s$, and the canonical filtration on $X_{N,s}^{[h]}$

$$0 \to \mathcal{C}_w^{\mathrm{conn}} \to \mathcal{C}_w \to \mathcal{C}_w^{\text{ét}} \to 0$$

has the property that $\mathcal{C}_w^{\text{ét}}$ is étale locally isomorphic to $(E_w/o_w)^{n-h}$.

We consider the formal completion $\mathcal{X}_N^{[h]} = \hat{X}_N|_{X_{N,s}^{[h]}}$ of X_N along $X_{N,s}^{[h]}$. Note that Hecke correspondences induce correspondences on $\mathcal{X}_N^{[h]}$.

Let P_h be the standard maximal parabolic subgroup of H_w which fixes the filtration $0 \subset o_w^h \subset o_w^n$.

Define $\mathcal{Y}_N^{[h]}$ to be the subspace of $\mathcal{X}_N^{[h]}$ where the canonical filtration of $\mathcal{C}_w[m_w^N]$ is compatible with the filtration $0 \subset (m_w^{-N}/o_w)^h \subset (m_w^{-N}/o_w)^n$ via the universal Drinfeld level structure $(m_w^{-N}/o_w)^n \to \mathcal{C}_w[m_w^N]$. $P_h(o_w/m_w^N)$ acts naturally on $\mathcal{Y}_N^{[h]}$.

We consider the pro-systems of formal schemes

$$\mathcal{X}_\infty^{[h]} = \varprojlim_N \mathcal{X}_N^{[h]} \quad \text{and} \quad \mathcal{Y}_\infty^{[h]} = \varprojlim_N \mathcal{Y}_N^{[h]}.$$

Note that $H_w(o_w)$ (resp. $P_h(o_w)$)-action on $\mathcal{X}_\infty^{[h]}$ (resp. $\mathcal{Y}_\infty^{[h]}$) is extended to $H_w(E_w)$ (resp. $P_h(E_w)$).

Then the canonical morphism[27]

$$H_w(E_w) \wedge^{P_h(E_w)} \mathcal{Y}_\infty^{[h]} \to \mathcal{X}_\infty^{[h]}$$

is in fact an isomorphism by a method of Boyer ([1], [20]), that is,

$$\mathcal{X}_\infty^{[h]} \simeq H_w(E_w) \wedge^{P_h(E_w)} \mathcal{Y}_\infty^{[h]}.$$

[27] For a group G and a right (resp. left) G-space X (resp. Y), $X \wedge^G Y$ denotes the contracted product, i.e., by viewing Y as a right G-space, $X \wedge^G Y = X \times Y/G$, where G acts diagonally on $X \times Y$.

Since the space itself is "parabolically induced from P_h", then for $\Lambda = \overline{\mathbb{F}}_\ell$, we have an isomorphism of admissible modules[28]

$$\varinjlim_N H^i_c(X^{[h]}_{N,\bar{s}}, R^j\psi(\Lambda)) \simeq \mathrm{Ind}_{P_h(E_w)}^{H_w(E_w)} \varinjlim_N H^i_c(Y^{[h]}_{N,\bar{s}}, R^j\psi(\Lambda)),$$

where $Y^{[h]}_N$ is the underlying scheme of $\mathcal{Y}^{[h]}_N$. In particular for $h < n$, this implies that the $H_v(E_w)$-representations occurring in proper support cohomologies $\varinjlim_N H^i_c(X^{[h]}_{N,\bar{s}}, R^j\psi(\Lambda))$ are induced from admissible representations of $P_h(E_w)$. Admissibility follows from the finiteness of nearby cycle cohomologies.

6.3. Vanishing theorem. We take a supercuspidal representation in the sense of Vignéras

$$\bar{\pi}_w : H_w(E_w) \to \mathrm{Aut}\, V_{\bar{\pi}_w}.$$

Here $V_{\bar{\pi}_w}$ is an $\overline{\mathbb{F}}_\ell$-vector space, that is, $\bar{\pi}_w$ is admissible, irreducible and cannot be obtained by a non-trivial parabolic induction. We assume that $\bar{\pi}_w^{K_w}$ is non-zero. $\bar{\pi}_w$ corresponds to a finite dimensional irreducible representation of $H(H_w(E_w),\ K_w)_{\overline{\mathbb{F}}_\ell}$.

Theorem 6.3 (Vanishing theorem for $\overline{\mathbb{F}}_\ell$-coefficients). *For a finite place w of E of residual characteristic p, assume that p is split in E_0, D is split at w, and $p \neq \ell$. For a compact open subgroup $K = K_w \cdot K^w$, assume moreover that K_w is contained in an Iwahori subgroup and K^w is sufficiently small. Then for any o_λ-local system $\mathcal{F}^K_{\nu,o_\lambda}$ on Sh_K corresponding to a finite dimensional representation ν of G, $\bar{\pi}_w$ does not appear as a subquotient of $H^i_B(\mathrm{Sh}_K,\ \mathcal{F}^K_{\nu,o_\lambda} \otimes_{o_\lambda} \bar{k}_\lambda)$ unless $i = n-1$.*

We give a sketch of the proof. We assume $\mathcal{F}^K_{\nu,o_\lambda} = o_\lambda$, since the general case is shown similarly. We set $\Lambda = \overline{\mathbb{F}}_\ell$. By the comparison theorem, the Betti cohomology is canonically isomorphic to the étale cohomology. We reduce the vanishing theorem to:

Lemma 6.4 (Localization principle). *For $X = \mathcal{S}h_K$, the kernel and the cokernel of the canonical map*

$$\begin{aligned} H^i_{\text{ét}}(X_{\bar{\eta}}, \Lambda) \simeq H^i_{\text{ét}}(X_{\bar{s}}, R\psi(\Lambda)) &\to H^i_{\text{ét}}(X^{[n]}_{\bar{s}}, R\psi(\Lambda)|_{X^{[n]}_{\bar{s}}}) \\ &= H^0_{\text{ét}}(X^{[n]}_{\bar{s}}, R^i\psi(\Lambda)|_{X^{[n]}_{\bar{s}}}) \end{aligned}$$

do not admit $\bar{\pi}_w$ as a subquotient, that is, the supercuspidal part of the global cohomology is isomorphic to nearby cycle fibers at "supersingular points".

To show the lemma, we assume that $K_w = K_{N,w}$ for simplicity[29]. For $\alpha \geq 1$, we denote $\mathcal{S}h_{K_\alpha}$ by X_α. Note that, for $\beta \geq \alpha \geq 1$, the $K_{\alpha,w}$-invariants $H^i_{\text{ét}}(X_{\beta,\bar{\eta}}, \Lambda)^{K_{\alpha,w}}$

[28] By using the regular base change theorem in étale cohomology to compare the algebraic and the formal situation.

[29] The general case follows from the perfect complex argument. Even the assumption on K_w can be weakened.

are $H^i_{\text{ét}}(X_{\alpha,\bar\eta}, \Lambda)$ since $K_{\alpha,w}/K_{\beta,w}$ has order prime to ℓ. The same is true for the nearby cycle fibers for X_α and X_β as the group action is free on the generic fiber, so it suffices to prove that the kernel and the cokernel of the homomorphism

$$\varinjlim_\alpha H^i_{\text{ét}}(X_{\alpha,\bar\eta}, \Lambda) \to \varinjlim_\alpha H^0_{\text{ét}}(X^{[n]}_{\alpha,\bar s}, R^i\psi(\Lambda)|_{X^{[n]}_{\alpha,\bar s}})$$

do not admit $\bar\pi_w$ as a subquotient. Note that both sides are admissible $H_w(E_w)$-representations. We work modulo the Serre subcategory of admissible representations generated by non-supercuspidal representations.

Then 6.4 follows from the result in 6.2, since the contribution of the proper support cohomologies of nearby cycle sheaves from $X^{[h]}_{N,s}$ for $h < n$ is parabolically induced from P_h.

We go back to the proof of 6.3. Since the nearby cycle is perverse,

$$R^i\psi(\Lambda)_{\bar x} = 0 \quad \text{for } i > n-1,$$

so we get

$$H^i_{\text{ét}}(X_{\bar\eta}, \Lambda) = 0 \quad \text{for } i > n-1$$

by the localization principle, modulo non-supercuspidal representations. By Poincaré duality on $X_{\bar\eta}$, we have the vanishing of the cohomologies of degree strictly less than $n-1$, since the contragradient of a supercuspidal representation is again supercuspidal.

Remark 6.5. The method of this proof also gives a vanishing theorem for $\mathbb{Q}_\ell$-sheaves. The result in this case is proved by Harris [19] by using Clozel's purity lemma and the base change lift to $D^{\text{op}\times}$.

Note that $X^{[n]}_{\bar s}$ consists of a single isogeny class preserving homogeneous polarization when we assume n is odd for simplicity (see [37], chapter 6). It is classified by an inner form G_- of G which is compact modulo the center at all infinite places. $G_-(\mathbb{Q}_p)$ is $\mathbb{Q}_p^\times \cdot \tilde D_w^{\text{op}\times} \cdot \prod_{u\in P_\wp, u\neq w} D_u^{\text{op}\times}$ for a division algebra $\tilde D_w$ over E_w of invariant $\frac{1}{n}$, and $G_-(\mathbb{A}_f^p) \simeq G(\mathbb{A}_f^p)$ holds, that is, G and G_- are locally isomorphic except at w and ι_0 as chosen in 6.1. Take a compact open subgroup $K_- = K_{-,w}\cdot K^w$ of $G_-(\mathbb{A}_f)$ with $K_{-,w} = o^{\text{op}\times}_{\tilde D_w}$ for a maximal order $o_{\tilde D_w}$ of $\tilde D_w$. Then the underlying space of $X^{[n]}_{\bar s}$ is regarded as a Hida variety $M_{K_-}(G_-, X_-)$ where X_- is a point. The key point for this identification is a theorem of Drinfeld that formal o_w-modules of dimension one and height n are unique over an algebraically closed field and that the endomorphism ring is isomorphic to $o_{\tilde D_w}$. So the canonical homomorphism

$$H^{n-1}_{\text{ét}}(X_{\bar\eta}, \Lambda) \to H^0_{\text{ét}}(X^{[n]}_{\bar s}, R^{n-1}\psi(\Lambda)|_{X^{[n]}_{\bar s}})$$

connects the middle dimensional cohomologies of Shimura and Hida varieties in characteristic p, though the sheaf $R^{n-1}\psi(\Lambda)|_{X^{[n]}_{\bar s}}$ may seem to be complicated at first glance. But the sheaf is described, for $\Lambda = \mathbb{Q}_\ell$, by Drinfeld's reciprocity

law as formulated by Carayol [6] and proved by Harris and Taylor in [20]. This is the geometric Jacquet–Langlands correspondence. More precisely, *the geometric Jacquet–Langlands isomorphism* for G and G_-.

6.4. Reciprocity law and Taylor–Wiles systems for unitary Shimura varieties. For an irreducible automorphic representation π of $U_D(\mathbb{A}_F)$, Clozel shows the existence of the base change lift $\mathrm{BC}(\pi)$ to $U_D(\mathbb{A}_E) \simeq D^{\mathrm{op}\times}(\mathbb{A}_E)$ [9]. For an automorphic representation π of G, $\mathrm{BC}(\pi) = (\psi, \Pi_D)$ is defined as a representation of $\mathbb{A}_E^\times \times D^{\mathrm{op}\times}(\mathbb{A}_E)$ [20].

$H^i_B(\mathrm{Sh}_K(G, X), \overline{\mathbb{Q}}_\ell)$ is decomposed by the irreducible automorphic $G(\mathbb{A})$-representations $\pi = \pi_f \otimes \pi_\infty$ such that π_∞ is cohomological for the trivial coefficient. We denote by $H^i_B(\mathrm{Sh}_K(G, X), \overline{\mathbb{Q}}_\ell)[\mathrm{cusp}]$ the part where Π_D corresponds to a cuspidal representation Π_{GL_n} of $\mathrm{GL}_n(\mathbb{A}_E)$ by the Jacquet–Langlands correspondence[30].

We use a Galois parameter $\rho\colon G_E \to (\mathrm{GL}_1 \times \mathrm{GL}_n)(\overline{\mathbb{Q}}_\ell)$ attached to ψ and Π_{GL_n} ([9], [20]). We denote by P_ρ the set of the isomorphism classes of the $G(\mathbb{A}_f)$-representation π'_f such that there is some π as above which has π'_f as the finite part and Galois parameter ρ.

Then the reciprocity law of Kottwitz, which is strengthened by Clozel, takes the form

$$H^{n-1}_{\text{ét}}(\mathrm{Sh}_K, \overline{\mathbb{Q}}_\ell)[\mathrm{cusp}] = \oplus_{\pi_f \in P_\rho} V_{\pi_f} \otimes \pi_f,$$

where $V^{\mathrm{ss}}_{\pi_f}$ is described by ρ [31].

Once the reciprocity is established, we are ready to construct a Taylor–Wiles system along the lines in §5, since we have the desired vanishing theorem in §6. Assume that $\bar\rho\colon G_E \to (\mathrm{GL}_1 \times \mathrm{GL}_n)(\overline{\mathbb{F}}_\ell)$ is obtained from a Galois parameter which appear in the reciprocity law by mod ℓ-reduction. We assume that D is split at $w \nmid \ell$ and that $\bar\rho|_{G_{E_w}}$ defines an absolutely irreducible representation on the standard representation of GL_n. With these assumptions, the vanishing assumption 5.1 is satisfied by using Theorem 6.3. We take a finite set Q of finite places of E so that Q is disjoint from all ramifications, $q_u \equiv 1 \mod \ell$, the Frobenius image $\bar\rho(\mathrm{Fr}_u)$ at u is a regular semi-simple element for all $u \in Q$. The compatibility of local and global Langlands correspondences plays an important role here.

Then the identification of a deformation ring R of $\bar\rho$ and the Hecke algebra T, as well as the freeness of the middle dimensional cohomology group M localized at the maximal ideal of the Hecke algebra, follows in the minimal case by the standard machinery in §4 under restrictive assumptions to define an appropriate deformation problem and to control the tangent space[32].

[30] Many technical conditions are omitted here. For example, at any place u of E, D_u is either split or a division algebra to be able to apply the global Jacquet–Langlands correspondence to $D^{\mathrm{op}\times}(\mathbb{A}_E)$.

[31] It is conjectured that V_{π_f} is an irreducible G_E-representation of dimension n. It seems likely that the assertion about the dimension is a consequence of the recent progress on the fundamental lemma by Laumon and Ngo.

[32] The basic assumptions are that: $\bar\rho$ has a large image, D is split at all places dividing ℓ and where $\bar\rho$ is ramified, and there is some π giving $\bar\rho$ such that π is minimally ramified and spherical at places dividing ℓ. More technical conditions are needed.

As is already mentioned in §5, Harris and Taylor already used Hida varieties associated to unitary groups which are compact at infinity to prove $R = T$ [33]. Thus our approach here does not give new information on the deformation ring, but it gives us hope that a further investigation might be possible for the middle dimensional cohomology of Shimura varieties along the lines sketched in §5.

7. Concluding remarks

Finally, the author would like to mention an application of the geometric Jacquet–Langlands correspondence to the theory of cohomological ℓ-adic automorphic forms of Emerton [16] and Urban which is closely related to the Coleman–Mazur construction of ℓ-adic families of automorphic forms for $\mathrm{GL}_{2,\mathbb{Q}}$ [10]. Assume $G_{\mathbb{Q}_\ell}$ is split with a split maximal torus $T_{\mathbb{Q}_\ell}$. We fix $K^\ell \subset G(\mathbb{A}_f^\ell)$. Consider the projective limit

$$\mathrm{Sh}_{K^\ell}(G, X) = \varprojlim_{K_\ell \subset G(\mathbb{Q}_\ell)} \mathrm{Sh}_{K_\ell \cdot K^\ell}(G, X).$$

By the vanishing theorem, under the supercuspidality assumption mod ℓ at a finite place $w \nmid \ell$, the ℓ-adic continuous complex $R\Gamma_{\mathrm{cont}}(\mathrm{Sh}_{K^\ell}(G, X), E_\lambda)$ reduces to one module H concentrated at the middle dimension. The Jacquet module of the space of locally analytic vectors in H is seen as the set of the global sections of a coherent sheaf $\mathcal{E}_{G,K^\ell}$ on the ℓ-adic rigid analytic torus $\mathcal{T}_\ell$ whose $\widehat{\overline{\mathbb{Q}}}_\ell$-valued point is $\mathrm{Hom}_{\mathrm{cont}}(T_{\mathbb{Q}_\ell}(\mathbb{Q}_\ell)/\overline{Z(F) \cap K^\ell}, \widehat{\overline{\mathbb{Q}}}_\ell^\times)$ (the weight space).

On the other hand, for the compact twist G_- in 6.3, the cohomological construction using Hida varieties gives a coherent sheaf $\mathcal{E}_{G_-,K_-^\ell}$ on $\mathcal{T}_\ell$. The geometric Jacquet–Langlands isomorphism in 6.3 suggests that $\mathcal{E}_{G,K^\ell}$ is described explicitly by $\mathcal{E}_{G_-,K_-^\ell}$, where $K^{\ell,w}$ is identified with $K_-^{\ell,w}$, preserving Hecke actions outside w. So the geometric Jacquet–Langlands correspondence, which preserves $\mathbb{Z}_\ell$-lattice structures (or even structures over $\mathbb{Z}_\ell/\ell^n$), realizes a correspondence between cohomological ℓ-adic automorphic forms on G and G_-. The existence of such a correspondence for GL_2 and the Coleman–Mazur construction was questioned by Buzzard ([3] for the construction on the definite quaternion side), and answered by Chenevier [8].

The author would like to close this article with the following speculation: the functoriality principle for reductive groups should be true even for cohomological ℓ-adic automorphic forms.

[33] After this paper was written, Taylor [53] has shown a potential automorphy for a large class of Galois representations by developing the techniques discussed in this article. As a consequence, he proved the Sato–Tate conjecture for elliptic curves quite generally.

References

[1] Boyer, P., Mauvaise réduction des variétés de Drinfeld et correspondance de Langlands locale. *Invent. Math.* **138** (3) (1999), 573–629.

[2] Breuil, C., Sur quelques représentations modulaires et p-adiques de $\mathrm{GL}_2(\mathbf{Q}_p)$. I. *Compositio Math.* **138** (2) (2003), 165–188.

[3] Buzzard, K.: On p-adic families of automorphic forms. *Modular curves and abelian varieties*, Progr. Math. 224, Birkhäuser, Basel 2004, 23–44.

[4] Carayol, H., Sur la mauvaise réduction des courbes de Shimura. *Compositio Math.* **59** (1986), 151–230.

[5] Carayol, H.,Sur les représentations p-adiques associées aux formes modulaires de Hilbert. *Ann. Sci. École Norm. Sup.* (4) **19** (1986), 409–468.

[6] Carayol, H.,Nonabelian Lubin-Tate theory. *Automorphic forms, Shimura varieties, and L-functions*, Vol. II, Perspect. Math. 11, Academic Press, Boston, MA, 1990, 15–39.

[7] Carayol, H.,Sur les représentations galoisiennes mod ℓ attachées aux formes modulaires. *Duke Math. J.* **59** (1989), 785–801.

[8] Chenevier, G.: Une correspondance de Jacquet-Langlands p-adique. *Duke Math. J.* **126** (1) (2005), 161–194.

[9] Clozel, L., Représentations galoisiennes associées aux représentations automorphes autoduales de $\mathrm{GL}(n)$. *Inst. Hautes Études Sci. Publ. Math.* **73** (1991), 97–145.

[10] Coleman, R., Mazur, B., The eigencurve. *Galois representations in arithmetic algebraic geometry* (Durham, 1996), London Math. Soc. Lecture Note Ser. 254, Cambridge University Press, Cambridge 1998, 1–113.

[11] Deligne, P., Travaux de Shimura. In *Séminaire Bourbaki*, 23ème année (1970/71), Exp. No. 389, Lecture Notes in Math. 244, Springer-Verlag, Berlin 1971, 123–165.

[12] Deligne, P., Variétés de Shimura: interprétation modulaire, et techniques de construction de modèles canoniques. In *Automorphic forms, representations and L-functions* (Corvallis, Ore., 1977), Part 2, Proc. Sympos. Pure Math. 33, Amer. Math. Soc., Providence, R.I., 1979, 247–289.

[13] Diamond, F., On deformation rings and Hecke rings. *Ann. of Math.* (2) **144** (1) (1996), 137–166.

[14] Diamond, F., The Taylor-Wiles construction and multiplicity one. *Invent. Math.* **128** (2) (1997), 379–391.

[15] Drinfeld, V. G., Elliptic modules. *Mat. Sb.* (*N.S.*) Ser. 94 **136** (1974), 594–627; English transl. *Math. USSR Sb.* **23** (1974), 561–592.

[16] Emerton, M., On the interpolation of eigenvalues attached to automorphic Hecke eigenforms. *Invent. Math.* **164** (2006), 1–84

[17] Fujiwara, K., Level optimization in the totally real case. Preprint, arXiv:math.NT/0602586.

[18] Fujiwara, K., Deformation rings and Hecke algebras in the totally real case. Preprint, arXiv:math.NT/0602606.

[19] Harris, M., Supercuspidal representations in the cohomology of Drinfeld upper half spaces; elaboration of Carayol's program. *Invent. Math.* **129** (1) (1997), 75–119.

[20] Harris, M., Taylor, R., *The geometry and cohomology of some simple Shimura varieties*. Ann. of Math. Stud. 151, Princeton University Press, Princeton, NJ, 2001.

[21] Hida, H., On p-adic Hecke algebras for GL_2 over totally real fields. *Ann. of Math.* **128** (1988), 295–384.

[22] Hida, H., Nearly ordinary Hecke algebras and Galois representations of several variables. In *Algebraic analysis, geometry, and number theory* (Baltimore, MD, 1988), Johns Hopkins University Press, Baltimore, MD, 1989, 115–134.

[23] Hida, H., Adjoint Selmer groups as Iwasawa modules. In *Proceedings of the Conference on p-adic Aspects of the Theory of Automorphic Representations* (Jerusalem, 1998); *Israel J. Math.* **120** (Part B) (2000), 361–427.

[24] Hida, H., The integral basis problem of Eichler. *Internat. Math. Res. Notices* **2005** (34) (2005), 2101–2122.

[25] Jacquet, H., Langlands, R. P., *Automorphic forms on* GL_2. Lecture Notes in Math. 114, Springer-Verlag, Berlin 1970.

[26] Jarvis, F., Mazur's principle for totally real fields of odd degree. *Compositio Math.* **116** (1999), 39–79.

[27] Jarvis, F., Level lowering for modular mod l representations over totally real fields. *Math. Ann.* **313** (1999), 141–160.

[28] Khare, C., On Serre's modularity conjecture for 2-dimensional mod p representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ unramified outside p. Preprint, arXiv:math.NT/0504080.

[29] Khare, C., Ramakrishna, R., Finiteness of Selmer groups and deformation rings. *Invent. Math.* **154** (1) (2003), 179–198.

[30] Khare, C., Wintenberger, J. P., On Serre's reciprocity conjecture for 2-dimensional mod p representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Preprint, arXiv:math.NT/0412076.

[31] Kisin, M., Modularity of potentially Barsotti-Tate representations and moduli of finite flat group schemes. Preprint.

[32] Kottwitz, R. E., Shimura varieties and λ-adic representations. In *Automorphic forms, Shimura varieties, and L-functions*, Vol. I, Perspect. Math. 10, Academic Press, Boston, MA, 1990, 161–209.

[33] Kottwitz, R. E., On the λ-adic representations associated to some simple Shimura varieties. Invent. Math. **108** (3) (1992), 653–665.

[34] Mokrane, A., Tilouine, J., Cohomology of Siegel varieties with p-adic integral coefficients and applications. Cohomology of Siegel varieties. *Astérisque* **280** (2002), 1–95.

[35] Ohta, M., On the zeta function of an abelian scheme over the Shimura curve. *Japan J. Math.* **9** (1983), 1–26.

[36] Rajaei, A., On the levels of mod l Hilbert modular forms. *J. Reine Angew. Math.* **537**, (2001), 33–65.

[37] Rapoport, M., Zink, Th., *Period spaces for p-divisible groups*. Ann. of Math. Stud. 141, Princeton University Press, Princeton, NJ, 1996.

[38] Ribet, K. A., Congruence relations between modular forms. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 503–514.

[39] Ribet, K. A., On modular representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ arising from modular forms. *Invent. Math.* **100** (1990), 431–476.

[40] Ribet, K. A., Report on mod ℓ representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. In *Motives* (Seattle, WA, 1991), Proc. Sympos. Pure Math. 55, Amer. Math. Soc., Providence, RI, 1994, 639–676.

[41] Saito, T., Modular forms and p-adic Hodge theory. *Invent. Math.* **129** (3) (1997), 607–620.

[42] Saito, T., Hilbert modular forms and p-adic Hodge theory. Preprint, 1999.

[43] Shimizu, H., On zeta functions of quaternion algebras. *Ann. of Math.* (2) **81** (1965), 166–193.

[44] Shimura, G., Correspondences modulaires et les fonctions zeta de courbes. *J. Math. Soc. Japan* **10** (1958), 1–28.

[45] Shimura, G., Number fields and zeta functions associated with discontinuous groups and algebraic varieties. In *Proceedings of the International Congress of Mathematicians* (Moscow, 1966), Izdat. "Mir", Moscow 1968, 290–299.

[46] Shimura, G., An ℓ-adic method in the theory of automorphic forms. In *Collected papers*, Vol. II: 1967–1977, Springer-Verlag, New York 2002, 237–272.

[47] Shimura, G., Construction of class fields and zeta functions of algebraic curves. *Ann. of Math.* **85** (1967), 58–159.

[48] Shimura, G., On canonical models of arithmetic quotient of bounded symmetric domains. *Ann. of Math.* **91** (1970), 144–222.

[49] Skinner, C. M., Wiles, A. J., Residually reducible representations and modular forms. *Inst. Hautes Études Sci. Publ. Math.* **89** (1999), 5–126.

[50] Skinner, C. M., Wiles, A. J., Base change and a problem of Serre. *Duke Math. J.* **107** (1) (2001), 15–25.

[51] Skinner, C. M., Wiles, A. J., Nearly ordinary deformations of irreducible residual representations. *Ann. Fac. Sci. Toulouse Math.* (6) **10** (1) (2001), 185–215.

[52] Taylor, R., On Galois representations associated to Hilbert modular forms. *Invent. Math.* **98** (1989), 265–280.

[53] Taylor, R., Automorphy for some ℓ-adic lifts of automorphic mod ℓ Galois representations, II. Preprint.

[54] Taylor, R., Yoshida, T., Compatibility of local and global Langlands correspondences. *J. Amer. Math. Soc.*, to appear.

[55] Taylor, R., Wiles, A. J., Ring theoretic properties of certain Hecke algebras. *Ann. of Math.* **141** (3) (1995), 553–572.

[56] Wiles, A. J., On ordinary λ-adic representations associated to Hilbert modular forms. *Invent. Math.* **94** (1988), 529–573.

[57] Wiles, A. J., Modular elliptic curves and Fermat's last theorem. *Ann. of Math.* **141** (1995), 443–551.

Graduate School of Mathematics, Nagoya University, Nagoya, Aichi, Japan

E-mail: fujiwara@math.nagoya-u.ac.jp

Generalising the Hardy–Littlewood method for primes

Ben Green*

Abstract. The Hardy–Littlewood method is a well-known technique in analytic number theory. Among its spectacular applications are Vinogradov's 1937 result that every sufficiently large odd number is a sum of three primes, and a related result of Chowla and Van der Corput giving an asymptotic for the number of 3-term progressions of primes, all less than N. This article surveys recent developments of the author and T. Tao, in which the Hardy–Littlewood method has been generalised to obtain, for example, an asymptotic for the number of 4-term arithmetic progressions of primes less than N.

Mathematics Subject Classification (2000). 11B25.

Keywords. Hardy–Littlewood method, prime numbers, arithmetic progressions, nilsequences.

1. Introduction

Godfrey Harold Hardy and John Edensor Littlewood wrote, in the 1920s, a famous series of papers *Some problems of "partitio numerorum"*. In these papers, whose content is elegantly surveyed by Vaughan [31], they developed techniques having their genesis in work of Hardy and Ramanujan on the partition function [18] to well-known questions in additive number theory such as Waring's problem and the Goldbach problem.

Papers III and V in the series, [16], [17], were devoted to the sequence of primes. In particular it was established on the assumption of the Generalised Riemann Hypothesis that every sufficiently large odd number is the sum of three primes. In 1937 Vinogradov [33] made a further substantial advance by removing the need for any unproved hypothesis.

The Hardy–Littlewood–Vinogradov method may be applied to give an asymptotic count for the number of solutions in primes p_i to any fixed linear equation

$$a_1 p_1 + \cdots + a_t p_t = b$$

in, say, the box $p_1, \ldots, p_t \leqslant N$, provided that at least 3 of the a_i are non-zero. This includes the three-primes result, and also the result that there are infinitely many

*This research was partially conducted during the period the author served as a Clay Research Fellow. He would like to express his sincere gratitude to the Clay Institute, and also to the Massachusetts Institute of Technology, where he was a Visiting Professor for the academic year 2005-06.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

triples of primes $p_1 < p_2 < p_3$ in arithmetic progression, due to Chowla [4] and van der Corput [29].

More generally the Hardy–Littlewood method may also be used to investigate systems such as $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$, where $\boldsymbol{A}$ is an $s \times t$ matrix with integer entries and, potentially, $s > 1$. A natural example of such a system is given by the $(k-2) \times k$ matrix

$$\boldsymbol{A} := \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ & & & & \dots & & & \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}, \tag{1}$$

in which case a solution to $\boldsymbol{A}\boldsymbol{p} = 0$ is just a k-term arithmetic progression of primes.

Here, unfortunately, the Hardy–Littlewood method falters in that it generally requires $t \geqslant 2s + 1$. In particular it cannot be used to handle progressions of length four or longer. There are certain special systems with fewer variables which *can* be handled. In this context we take the opportunity to mention a beautiful result of Balog [2], where it is shown that for any m there are distinct primes $p_1 < \cdots < p_m$ such that each number $\frac{1}{2}(p_i + p_j)$ is also prime, or in other words that the system

$$\begin{aligned} p_1 + p_2 &= 2p_{12} \\ &\vdots \\ p_{m-1} + p_m &= 2p_{m-1,m} \end{aligned} \tag{2}$$

has a solution in primes $p_1, \dots, p_m, p_{12}, \dots, p_{m-1,m}$. There is also a result of Heath-Brown [19], in which it is established that there are infinitely many four-term progressions in which three members are prime and the fourth is either a prime or a product of two primes.

The survey of Kumchev and Tolev [25] gives a detailed account of applications of the Hardy–Littlewood method to additive prime number theory.

The aim of this survey is to give an overview of recent joint work with Terence Tao [13], [14], [15]. Our aim, which has been partially successful, is to extend the Hardy–Littlewood method so that it is capable of handling a more-or-less arbitrary system $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$, subject to the proviso that we do not expect to be able to handle any system which secretly encodes a "binary" problem such as Goldbach or Twin Primes.

This is a large and somewhat technical body of work. Perhaps my main aim here is to give a guide to our work so far, pointing out ways in which the various papers fit together, and future directions we plan to take. A subsidiary aim is to focus as far as possible on key concepts, rather than on details. Of course, one would normally aim to do this in a survey article. However in our case we expect that many of these details will be substantially cleaned up in future incarnations of the theory, whilst the key concepts ought to remain more-or-less as they are.

I will say rather little about our paper [12] establishing that there are arbitrarily long arithmetic progressions of primes. Whilst there is considerable overlap between

that paper and the ideas we discuss here, those methods were somewhat "soft" whereas the flavour of our more recent work is distinctly "hard". We refer the reader to the survey of Tao in Volume II of these Proceedings, and also to the surveys [11], [23], [27], [28].

To conclude this introduction let me remark that the reader should not be under the impression that the Hardy–Littlewood method only applies to linear equations in primes, or even that this is the most popular application of the method. There has, for example, been a huge amount done on the circle of questions surrounding Waring's problem. For a survey see [32]. More generally there are many spectacular results where variants of the method are used to locate integer points on quite general varieties, provided of course that there are sufficiently many variables. The reader may consult Wooley's survey [34] for more information on this.

2. The Hardy–Littlewood heuristic

We have stated our interest in systems of linear equations in primes. While we are still somewhat lacking in theoretical results, there are heuristics which predict what answers we should expect in more-or-less any situation.

It is natural, when working with primes, to introduce the von Mangoldt function $\Lambda\colon \mathbb{N} \to \mathbb{R}_{\geqslant 0}$, defined by

$$\Lambda(n) := \begin{cases} \log p & \text{if } n = p^k \text{ is a prime power,} \\ 0 & \text{otherwise.} \end{cases}$$

The prime powers with $k \geqslant 2$ make a negligible contribution to any additive expression involving Λ. Thus, for example, the prime number theorem is equivalent to the statement that

$$\mathbb{E}_{n \leqslant N} \Lambda(n) = 1 + o(1).$$

Here we have used the very convenient notation of expectation from probability theory, setting $\mathbb{E}_{x \in X} := |X|^{-1} \sum_{x \in X}$ for any set X. The notation $o(1)$ refers to a quantity which tends to zero as $N \to \infty$.

We now discuss a version of the Hardy–Littlewood heuristic for systems of linear equations in primes. Here, and for the rest of the article, we restrict attention to homogeneous systems for simplicity of exposition.

Conjecture 2.1 (Hardy–Littlewood). Let $\boldsymbol{A}$ be a fixed $s \times t$ matrix with integer entries and such that there is at least one non-zero solution to $\boldsymbol{A}\boldsymbol{x} = 0$ with $x_1, \dots, x_t \geqslant 0$. Then

$$\mathbb{E}_{\substack{x_1,\dots,x_t \leqslant N \\ \boldsymbol{A}\boldsymbol{x}=0}} \Lambda(x_1) \dots \Lambda(x_t) = \mathfrak{S}(A)(1 + o(1))$$

as $N \to \infty$, where the *singular series* $\mathfrak{S}(A)$ is equal to a product of local factors $\prod_p \alpha_p$, where

$$\alpha_p := \left(\frac{p}{p-1}\right)^t \lim_{M\to\infty} \frac{\mathbb{P}(\boldsymbol{A}\boldsymbol{x} = 0, (x_1, p) = \cdots = (x_t, p) = 1 | \boldsymbol{x} \in [-M, M]^t)}{\mathbb{P}(\boldsymbol{A}\boldsymbol{x} = 0 | \boldsymbol{x} \in [-M, M]^t)}.$$

The singular series reflects "local obstructions" to having solutions to $\boldsymbol{A}\boldsymbol{x} = 0$ in primes; in the simple example $\boldsymbol{A} = \begin{pmatrix} 1 & 9 & -27 \end{pmatrix}$, where the associated equation $p_1 + 9p_2 - 27p_3 = 0$ has no solutions, one has $\alpha_3 = 0$. A more elegant formulation of the conjecture would include a "local obstruction at ∞" α_∞, in exchange for removing the hypothesis on $\boldsymbol{A}$.

Chowla and van der Corput's results concerning three-term progressions of primes confirm the prediction Conjecture 2.1 for the matrix $\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}$. From this it is easy to derive an *asymptotic* for the number of triples (p_1, p_2, p_3), $p_1 < p_2 < p_3 \leqslant N$, of primes in arithmetic progression.

Theorem 2.2 (Chowla, van der Corput, [4], [29]). *The number of triples of primes (p_1, p_2, p_3), $p_1 < p_2 < p_3 \leqslant N$, in arithmetic progression is*

$$\mathfrak{S}_3 N^2 \log^{-3} N(1 + o(1)),$$

where

$$\mathfrak{S}_3 := \frac{1}{2} \prod_{p \geqslant 3} \left(1 - \frac{1}{(p-1)^2}\right) \approx 0.3301.$$

The singular series $\mathfrak{S}_3$ is equal to $\frac{1}{4}\mathfrak{S}(\boldsymbol{A})$, where $\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}$, and is also half the twin prime constant.

Certain systems $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$ should be thought of as very difficult indeed, since their understanding implies an understanding of a binary problem such as the Goldbach or twin prime problem. If $\boldsymbol{A}$ has the property that every non-zero vector in its row span (over $\mathbb{Q}$) has at least three non-zero entries then there is no such reason to believe that it should be fantastically hard to solve.

Definition 2.3 (Non-degenerate systems). Suppose that s, t are positive integers with $t \geqslant s + 2$. We say that an $s \times t$ matrix $\boldsymbol{A}$ with integer entries is *non-degenerate* if it has rank s, and if every non-zero vector in its row span (over $\mathbb{Q}$) has at least three non-zero entries.

The reader may care to check that the system (1) defining a progression of length k is non-degenerate.

Our eventual goal is to prove Conjecture 2.1 for all non-degenerate systems. This goal may be subdivided into subgoals according to the value of s.

Conjecture 2.4 (Asymptotics for s simultaneous equations). Fix a value of $s \geqslant 1$ and suppose that $t \geqslant s + 2$ and that $\boldsymbol{A}$ is a non-degenerate $s \times t$ matrix. Then Conjecture 2.1 holds for the system $\boldsymbol{A}\boldsymbol{p} = 0$.

One can also formulate an appropriate conjecture for non-homogeneous systems $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$, and one would not expect to encounter significant extra difficulties in proving it. One might also try to count prime solutions to $\boldsymbol{A}\boldsymbol{p} = 0$ in which the primes p_i are subject to different constraints $p_i \leqslant N_i$, or perhaps are constrained to lie in a fixed arithmetic progression $p_i \equiv a_i \pmod{q_i}$. One would expect all of these extensions to be relatively straightforward.

The classical Hardy–Littlewood method can handle the case $s = 1$ of Conjecture 2.4. Our new developments have led to a solution of the case $s = 2$. In particular we can obtain an asymptotic for the number of 4-term arithmetic progressions of primes, all less than N:

Theorem 2.5 (Green–Tao [15]). *The number of quadruples of primes* (p_1, p_2, p_3, p_4), $p_1 < p_2 < p_3 < p_4 \leqslant N$, *in arithmetic progression is*

$$\mathfrak{S}_4 N^2 \log^{-4} N(1 + o(1)),$$

where

$$\mathfrak{S}_4 := \frac{3}{4} \prod_{p \geqslant 5} \left(1 - \frac{3p-1}{(p-1)^3}\right) \approx 0.4764.$$

3. The Hardy–Littlewood method for primes

The aim of this section is to describe the Hardy–Littlewood method as it would normally be applied to linear equations in primes. We will sketch the proof of Theorem 2.2, the asymptotic for the number of 3-term progressions of primes. This is equivalent to the $s = 1$ case of Conjecture 2.4 for the specific matrix $\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}$. Very similar means may be used to handle the general case $s = 1$ of that conjecture.

The Hardy–Littlewood method is, first and foremost, a method of harmonic analysis. The primes are studied by introducing the *exponential sum* (a kind of Fourier transform)

$$S(\theta) := \mathbb{E}_{n \leqslant N} \Lambda(n) e(\theta n)$$

for $\theta \in \mathbb{R}/\mathbb{Z}$, where $e(\alpha) := e^{2\pi i \alpha}$. It is the appearance of the circle $\mathbb{R}/\mathbb{Z}$ here which gives the Hardy–Littlewood method its alternative name. Now it is easy to check that

$$\mathbb{E}_{x_1, x_2, x_3 \leqslant N} \Lambda(x_1)\Lambda(x_2)\Lambda(x_3) 1_{x_1 - 2x_2 + x_3 = 0} = \int_0^1 S(\theta)^2 S(-2\theta)\, d\theta.$$

whence

$$\mathbb{E}_{\substack{x_1, x_2, x_3 \leqslant N \\ x_1 - 2x_2 + x_3 = 0}} \Lambda(x_1)\Lambda(x_2)\Lambda(x_3) = (2N + O(1)) \int_0^1 S(\theta)^2 S(-2\theta)\, d\theta. \qquad (3)$$

The method consists of gathering information about $S(\theta)$, and then using this formula to infer an asymptotic for the left-hand side.

The process of gathering information about $S(\theta)$ leads us to another key feature of the Hardy–Littlewood method: the realisation that one must split the set of θ into two classes, the *major arcs* $\mathfrak{M}$ in which $\theta \approx a/q$ for some small q and the *minor arcs* $\mathfrak{m} := [0, 1) \setminus \mathfrak{M}$. To see why, let us attempt some simple evaluations. First of all we note that

$$S(0) := \mathbb{E}_{n \leqslant N} \Lambda(n) = 1 + o(1),$$

this being equivalent to the prime number theorem. To evaluate $S(1/2)$, observe that almost all of the support of Λ is on odd numbers n, for which $e(n/2) = -1$. Thus

$$S(1/2) := \mathbb{E}_{n \leqslant N} \Lambda(n) e(n/2) = -1 + o(1).$$

The evaluation of $S(1/3)$ is a little more subtle. Most of the support of Λ is on n not divisible by 3, and for those n the character $e(n/3)$ takes two values according as $n \equiv 1 \,(\mathrm{mod}\ 3)$ or $n \equiv 2 \,(\mathrm{mod}\ 3)$. We have

$$\begin{aligned} S(1/3) &= e(1/3)\mathbb{E}_{n \leqslant N} 1_{n \equiv 1 \,(\mathrm{mod}\ 3)} \Lambda(n) + e(2/3)\mathbb{E}_{n \leqslant N} 1_{n \equiv 2 \,(\mathrm{mod}\ 3)} \Lambda(n) + o(1) \\ &= -1/2 + o(1), \end{aligned}$$

this being a consequence of the fact that the primes are asymptotically equally divided between the congruence classes $1 \,(\mathrm{mod}\ 3)$ and $2 \,(\mathrm{mod}\ 3)$.

In similar fashion one can get an estimate for $S(a/q)$ for small q, and indeed for $S(a/q+\eta)$ for sufficiently small η, if one uses the prime number theorem in arithmetic progressions. The set of such θ is called the *major arcs* and is denoted $\mathfrak{M}$. (The notion of "small q" might be $q \leqslant \log^A N$, for some fixed A. The notion of "small η" might be $|\eta| \leqslant \log^A N/qN$. The flexibility allowed here depends on what type of prime number theorem along arithmetic progressions one is assuming. Unconditionally, the best such theorem is due to Siegel and Walfisz and it is this theorem which leads to these bounds on q and $|\eta|$.)

Suppose by contrast that $\theta \notin \mathfrak{M}$, that is to say θ is not close to a/q with q small. We say that $\theta \in \mathfrak{m}$, the *minor arcs*. It is hard to imagine that in the sum

$$S(\sqrt{2} - 1) = \mathbb{E}_{n \leqslant N} \Lambda(n) e(n\sqrt{2}) \tag{4}$$

the phases $e(n\sqrt{2})$ could conspire with $\Lambda(n)$ to prevent cancellation. It turns out that indeed there *is* substantial cancellation in this sum. This was first proved by Vinogradov, and nowadays it is most readily established using an identity of Vaughan [30], which allows one to decompose (4) into three further sums which are amenable to estimation. We will discuss a variant of this method in §5. For the particular value $\theta = \sqrt{2} - 1$, and for other highly irrational values, one can obtain an estimate of the shape $|S(\theta)| \ll N^{-c}$ for some $c > 0$, which is quite remarkable since applying the best-known error term in the prime number theorem only allows one to estimate $S(0)$ with the much larger error $O(\exp(-C_\varepsilon \log^{3/5-\varepsilon} N))$. By defining parameters suitably (that is by taking a suitable value of the constant A in the precise definition

of $\mathfrak{M}$), one can arrange that $S(\theta)$ is always very small indeed on the minor arcs $\mathfrak{m}$, say

$$\sup_{\theta\in\mathfrak{m}} |S(\theta)| \ll \log^{-10} N. \tag{5}$$

Recall now the formula (3). Splitting the integral into that over $\mathfrak{M}$ and that over $\mathfrak{m}$, we see from Parseval's identity that

$$\left| \int_{\mathfrak{m}} S(\theta)^2 S(-2\theta)\, d\theta \right| \leqslant \sup_{\theta\in\mathfrak{m}} |S(\theta)| \int_0^1 |S(\theta)|^2\, d\theta \ll \frac{\log^{-9} N}{N}. \tag{6}$$

Thus in the effort to establish Theorem 2.2 the contribution from the minor arcs $\mathfrak{m}$ may essentially be ignored. The proof of that theorem is now reduced to showing that

$$\int_{\mathfrak{M}} S(\theta)^2 S(-2\theta)\, d\theta = (1+o(1))\frac{1}{N}\prod_{p\geqslant 3}\Big(1-\frac{1}{(p-1)^2}\Big).$$

Since one has asymptotic formulæ for $S(\theta)$ (and $S(-2\theta)$) on $\mathfrak{M}$, this is essentially just a computation, albeit not a particularly straightforward one.

It is instructive to look for the point in the above argument where we used the fact that $\boldsymbol{A}$ was non-degenerate, that is to say that our problem had at least three variables. Why can we not use the same ideas to solve the twin prime or Goldbach problems? The answer lies in the bound (6). In the twin prime problem we would be looking to bound

$$\left| \int_{\theta\in\mathfrak{m}} |S(\theta)|^2 e(2\theta) \right|,$$

and the only obvious means of doing this is via an inequality of the form

$$\left| \int_{\theta\in\mathfrak{m}} |S(\theta)|^2 e(2\theta) \right| \leqslant \sup_{\theta\in\mathfrak{m}} |S(\theta)|^c \int_0^1 |S(\theta)|^{2-c}\, d\theta.$$

Now, however, Parseval's identity does not permit one to place a bound on

$$\int_0^1 |S(\theta)|^{2-c}\, d\theta.$$

Indeed this whole endeavour is rather futile since heuristics predict that the minor arcs actually make a significant contribution to the asymptotic for twin primes.

An attempt to count 4-term progressions in primes via the circle method is beset by difficulties of a similar kind.

4. Exponential sums with Möbius

The presentation in the next two sections (and in our papers) is influenced by that in the beautiful book of Iwaniec and Kowalski [21].

In the previous section we described what is more-or-less the standard approach to solving linear equations in primes using the Hardy–Littlewood method. In [21, Ch. 19] one may find a very elegant variant in which the Möbius function μ is made to play a prominent rôle. As we saw above the behaviour of the exponential sum $S(\theta)$ was a little complicated to describe, depending as it does on how close to a rational θ is. By contrast the exponential sum

$$M(\theta) := \mathbb{E}_{n\leqslant N}\mu(n)e(\theta n)$$

has a very simple behaviour, as the following result of Davenport shows.

Proposition 4.1 (Davenport's bound). *We have the estimate*

$$|M(\theta)| \ll_A \log^{-A} N$$

uniformly in $\theta \in [0, 1)$ *for any* $A > 0$.

In fact on the GRH Baker and Harman [1] obtain the superior bound $|M(\theta)| \ll N^{-3/4+\varepsilon}$. By analogy with results of Salem and Zygmund [26] concerning random trigonometric series one might guess that the truth is that $\sup_{\theta\in[0,1)} |M(\theta)| \sim c\sqrt{\log N/N}$. This is far from known even on GRH; so far as I am aware no *lower* bound of the form $\sup_{\theta\in[0,1)} |M(\theta)|\sqrt{N} \to \infty$ is known.

Although Davenport's result is easy to describe its proof has the same ingredients as used in the analysis of $S(\theta)$. One must again divide $\mathbb{R}/\mathbb{Z}$ into major and minor arcs. On the major arcs one must once more use information equivalent to a prime number theorem along arithmetic progressions, that is to say information on the zeros of L-functions $L(s, \chi)$ close to the line $\Re s = 1$. On the minor arcs one uses an appropriate version of Vaughan's identity. One of the attractions of working with Möbius is that this identity takes a particularly simple form (see [21, Ch. 13] or [14]).

We offer a rough sketch of how Proposition 4.1 may be used as the main ingredient in a proof of Theorem 2.2, referring the reader to [21, Ch. 19] for the details. The key point is that one has the identity

$$\Lambda(n) = \sum_{d|n} \mu(d) \log(n/d).$$

One splits the sum over d into the ranges $d \leqslant N^{1/10}$ and $d > N^{1/10}$ (say), obtaining a decomposition $\Lambda = \Lambda^\sharp + \Lambda^\flat$. One has

$$S^\flat(\theta) := \mathbb{E}_{n\leqslant N}\Lambda^\flat(n)e(n\theta) = \sum_{d\leqslant N^{1/10}} \log d \sum_{N^{1/10}\leqslant k\leqslant N/d} \mu(k)e(\theta kd),$$

from which it follows easily using Davenport's bound that

$$S^\flat(\theta) \ll_A \log^{-A} N \tag{7}$$

uniformly in $\theta \in [0, 1)$.

One may then write the expression

$$\mathbb{E}_{\substack{x_1,x_2,x_3 \leqslant N \\ x_1-2x_2+x_3=0}} \Lambda(x_1)\Lambda(x_2)\Lambda(x_3)$$

as a sum of eight terms using the splitting $\Lambda = \Lambda^\sharp + \Lambda^\flat$. The basic idea is now that the main term $\prod_p \alpha_p$ in Theorem 2.2 comes from the term with three copies of $\Lambda^\sharp$, whilst the other 7 terms (each of which contains at least one $\Lambda^\flat$) provide a negligible contribution in view of (7) and simple variants of the formula (3).

We have extolled the virtues of the Möbius function by pointing to the aesthetic qualities of Davenport's bound. A more persuasive argument for focussing on it is the following basic metaprinciple of analytic number theory:

Principle (Möbius randomness law). *The Möbius function is highly orthogonal to any "reasonable" bounded function $f : \mathbb{N} \to \mathbb{C}$. That is to say*

$$\mathbb{E}_{n \leqslant N} \mu(n) f(n) = o(1),$$

and usually one would in fact expect

$$\mathbb{E}_{n \leqslant N} \mu(n) f(n) \ll N^{-1/2+\varepsilon}. \tag{8}$$

In the category "reasonable" in this context one would certainly include polynomials phases and other somewhat continuous objects, but one should exclude functions f which are closely related to the primes ($f = \mu$ and $f = \Lambda$, for example, are clearly not orthogonal to Möbius).

At a finer level than is relevant to our work, the Möbius randomness law is more reliable than other heuristics that one might formulate, for example concerning Λ. In [22] it is shown that

$$\mathbb{E}_{n \leqslant N} \Lambda(n)\lambda(n)e(-2\sqrt{n}) \sim cN^{-1/4},$$

where $\lambda(n) := n^{-11/2}\tau(n)$ is a normalised version of Ramanujan's τ-function. One could hardly be called naïve for expecting square root cancellation here.

5. Proving the Möbius randomness law

In the last section we mentioned a principle, the *Möbius randomness law,* which is very useful as a guiding principle in analytic number theory. Unfortunately it is not possible to prove the strong version (8) of the principle in any case – even when $f(n) \equiv 1$ it is equivalent to the Riemann hypothesis.

It is, however, possible to prove weaker estimates of the form

$$\mathbb{E}_{n \leqslant N} \mu(n) f(n) \ll_A \log^{-A} N, \tag{9}$$

for arbitrary $A > 0$, for a wide variety of functions f. Davenport's bound is precisely this result when $f(n) = e(\theta n)$ (and, furthermore, this result is uniform in θ). Similar statements are also known for polynomial phases and for Dirichlet characters (uniformly over all characters of a fixed conductor).

Now when it comes to proving an estimate of the form (9), one should think of there being two different classes of behaviour for f. In the first class are those f which are in a vague sense multiplicative, or linear combinations of a few multiplicative functions. Then the behaviour of $\mathbb{E}_{n\leqslant N}\mu(n)f(n)$ can be intimately connected with the zeros of L-functions. One has, for example, the formula

$$\sum_{n=1}^{\infty}\mu(n)\chi(n)n^{-s} = \frac{1}{L(s,\chi)}$$

for any fixed Dirichlet character χ. By the standard contour integration technique (Perron's formula) of analytic number theory one sees that $\mathbb{E}_{n\leqslant N}\mu(n)\chi(n)$ is small provided that $L(s,\chi)$ does not have zeros close to $\Re s = 1$. (In fact, as reported on [21, p. 124], there are complications caused by possible multiple zeros of L, and it is better to work first with the sum $\mathbb{E}_{n\leqslant N}\Lambda(n)\chi(n)$ of χ over primes.)

The need to consider zeros of L-functions can also be felt when considering *additive* characters $e(an/q)$, for relatively small q. Indeed any Dirichlet character to the modulus q may be expressed as a linear combination of such characters. Conversely any additive character $e(an/q)$ may be written as a linear combination of Dirichlet characters to moduli dividing q by using Gauss sums. By applying Siegel's theorem, which gives the best unconditional information concerning the location of zeros of $L(s,\chi)$ near to $\Re s = 1$, one obtains for any A the estimate

$$\mathbb{E}_{n\leqslant N}\mu(n)e(an/q) \ll_A \log^{-A} N,$$

uniformly for $q \leqslant \log^A N$. By partial summation the same estimate holds when a/q is replaced by $\theta = a/q + \eta$ for suitably small η, that is to say for all θ which lie in the set $\mathfrak{M}$ of major arcs.

We turn now to a completely different technique for bounding $\mathbb{E}_{n\leqslant N}\mu(n)f(n)$. Remarkably this is at its most effective when the previous technique fails, that is to say when f is somehow *far* from multiplicative.

Proposition 5.1 (Type I and II sums control sums with Möbius). *Let $f:\mathbb{N}\to\mathbb{C}$ be a function with $\|f\|_\infty \leqslant 1$, and suppose that the following two estimates hold.*

1. (Type I sums are small) *For all $D \leqslant N^{2/3}$, and for all sequences $(a_d)_{d=D}^{2D}$ with $\|a\|_{l^2[D,2D)} = 1$, we have*

$$\Big|\sum_{d=D}^{2D}\sum_{1\leqslant w<N/d} a_d f(wd)\Big| \ll_A N(\log N)^{-A-3}. \tag{10}$$

2. (Type II sums are small) *For all* D, W, $N^{1/3} \leqslant D \leqslant N^{2/3}$, $N^{1/3} \leqslant W \leqslant N/D$ *and all choices of complex sequences* $(a_d)_{d=D}^{2D}$, $(b_w)_{w=W}^{2W}$ *with* $\|a\|_{l^2[D,2D)} = \|b\|_{l^2[W,2W)} = 1$, *we have*

$$\Big|\sum_{d=D}^{2D}\sum_{W\leqslant w\leqslant 2W} a_d b_w f(wd)\Big| \ll_A N(\log N)^{-A-5}. \tag{11}$$

Then

$$\mathbb{E}_{n\leqslant N}\mu(n)f(n) \ll_A \log^{-A} N. \tag{12}$$

The reader may find a proof of this statement in [14, Ch. 6]. It is proved by decomposing the Möbius function into two parts using an identity of Vaughan [30]. When one multiplies by $f(n)$ and sums, one of these parts leads to Type I sums and the other to Type II sums. Note that there is considerable flexibility in arranging the ranges of D in which Type I and II estimates are required, but it is not important to have such flexibility in our arguments.

The statement of Proposition 5.1 may look complicated. What has been achieved, however, is the elimination of μ. Strictly speaking, one actually only *needs* Type I and II estimates for some rather specific choices of coefficients a_d, b_w whose definition involves μ. The important realisation is that it is best to forget about the precise forms of these coefficients, the general expressions (10) and (11) laying bare the important underlying information required of f.

Note that if f is close to multiplicative then there is no hope of obtaining enough cancellation in Type II sums to make use of Proposition 5.1. If f is actually completely multiplicative, for example, one may take $a_d = \overline{f(d)}$ and $b_w = \overline{f(w)}$ and there is manifestly no cancellation at all in (11). If this is not the case, however, then very often it *is* possible to verify the bounds (10) and (11). An example of this is a linear phase $e(\theta n)$ where θ lies in the minor arcs $\mathfrak{m}$, that is to say θ is not close to a/q with q small. By verifying these two estimates for such θ, one has from (12) that Davenport's bound holds when $\theta \in \mathfrak{m}$. This completes the proof of Davenport's bound, since the major arcs $\mathfrak{M}$ have already been handled using L-function technology.

To see how this is usually achieved in practice we refer the reader to [5, Ch. 24]. There the reader will see that a key device is the Cauchy–Schwarz inequality, which allows one to eliminate the arbitrary coefficients a_d, b_w.

In [14] there is also a discussion of this result. Although logically equivalent, this discussion takes a point of view which turns out to be invaluable when dealing with more complicated situations. Taking $f(n) = e(\theta n)$ in Proposition 5.1, we suppose that either (10) or (11) does not hold, that is to say that either a Type I or a Type II sum is large. We then *deduce* that θ must be close to a rational with small denominator, that is to say θ must be major arc. This *inverse* approach to bounding sums with Möbius means that there is no need to make an *a priori* definition of what a "major" or "minor" object is. In situations to be discussed later this helps enormously.

6. The insufficiency of harmonic analysis

What did we mean when we stated that the Hardy–Littlewood method was a method of harmonic analysis? In §3 we saw that there is a formula, (3), which expresses the number of 3-term progressions in a set (such as the primes) in terms of the exponential sum over that set. The following proposition is an easy consequence of a slightly generalised version of that formula:

Proposition 6.1. *Suppose that* $f_1, f_2, f_3 \colon [N] \to [-1, 1]$ *are three functions and that*

$$|\mathbb{E}_{\substack{x_1,x_2,x_3 \\ x_1-2x_2+x_3=0}} f_1(x_1) f_2(x_2) f_3(x_3)| \geqslant \delta.$$

Then for any $i = 1, 2, 3$ *we have*

$$\sup_{\theta\in[0,1)} |\mathbb{E}_{n\leqslant N} f_i(n)e(n\theta)| \geqslant (1 + o(1))\delta/2. \tag{13}$$

We think of this as a statement the effect that the linear exponentials $e(n\theta)$ form a *characteristic system* for the linear equation $x_1-2x_2+x_3 = 0$. It follows immediately from Proposition 6.1 and Davenport's bound that Möbius exhibits cancellation along 3-term APs, in the sense that

$$\mathbb{E}_{\substack{x_1,x_2,x_3 \\ x_1-2x_2+x_3=0}} \mu(x_1)\mu(x_2)\mu(x_3) \ll_A \log^{-A} N.$$

Proposition 6.1 is also useful for counting progressions in sets $A \subseteq [N]$, in which context one would take various of the f_i to equal the *balanced function* $f_A := 1_A - \alpha$ of A, where $\alpha := |A|/N$. It is easy to deduce from Proposition 6.1 the following variant, which covers this situation.

Proposition 6.2. *Suppose that* $A \subseteq [N]$ *is a set with* $|A| = \alpha N$ *and that*

$$|\mathbb{E}_{\substack{x_1,x_2,x_3 \\ x_1-2x_2+x_3=0}} 1_A(x_1)1_A(x_2)1_A(x_3) - \alpha^3| \geqslant \delta.$$

Write

$$f_A := 1_A - \alpha$$

for the balanced function of A*. Then we have*

$$\sup_{\theta\in[0,1)} |\mathbb{E}_{n\leqslant N} f_A(n)e(n\theta)| \geqslant (1 + o(1))\delta/14. \tag{14}$$

If a function f correlates with a linear exponential as in (13) or (14) then we sometimes say that f has *linear bias*.

In this section we give examples which show that the linear exponentials do not form a characteristic system for the pair of equations $x_1-2x_2+x_3 = x_2-2x_3+x_4 = 0$ defining a four-term progression. These examples show, in a strong sense, that the Hardy–Littlewood method in its traditional form *cannot* be used to study 4-term

progressions. An interesting feature of these two examples is that they were both essentially discovered by Furstenberg and Weiss [6] in the context of ergodic theory. Much of our work is paralleled in, and in fact motivated by, the work of the ergodic theory community. See the lecture by Tao in Volume II of these proceedings, or the elegant surveys of Kra [23], [24] for more discussion and references. The examples were rediscovered, in the finite setting, by Gowers [8], [10] in his work on Szemerédi's theorem.

Example 6.1 (Quadratic and generalised quadratic behaviour). Let $\alpha > 0$ be a small, fixed, real number, and define the following sets. Let A_1 be defined by

$$A_1 := \{x \in [N] : \{x^2\sqrt{2}\} \in [-\alpha/2, \alpha/2]\}$$

(here, $\{t\}$ denotes the fractional part of t, and lies in $(-1/2, 1/2]$). Define also

$$A_2 := \{x \in [N] : \{x\sqrt{2}\{x\sqrt{3}\}\} \in [-\alpha/2, \alpha/2]\}.$$

Now it can be shown (not altogether straightforwardly) that $|A_1|, |A_2| \approx \alpha N$, and furthermore that

$$\sup_{\theta\in[0,1)} |\mathbb{E}_{n\leqslant N} f_A(n)e(n\theta)| \ll N^{-c}$$

for $i = 1, 2$. Thus neither of the sets A_1, A_2 has linear bias in a rather strong sense. If the analogue of Proposition 6.2 were true for four term progressions, then, one would expect both A_1 and A_2 to have approximately $\alpha^4 N^2/6$ four-term progressions.

The set A_1, however, has considerably more 4-term APs that this in view of the identity

$$x^2 - 3(x + d)^2 + 3(x + 2d)^3 - (x + 3d)^2 = 0. \tag{15}$$

This means that if $x, x + d, x + 2d \in A_1$ then

$$\{(x + 3d)^2\sqrt{2}\} \in [-7\alpha/2, 7\alpha/2],$$

which would suggest that $x + 3d \in A_1$ with probability $\gg 1$. In fact one can show using harmonic analysis that (15) is the only relevant constraint in the sense that

$$\begin{aligned}\mathbb{P}(x + 3d \in A_1 | x, x + d, x + 2d \in A_1)& \\ \approx \mathbb{P}(y_1 - 3y_2 + 3y_3 \in [-1, 1] | y_1, y_2, y_3 \in [-1, 1]) &= 8/27.\end{aligned}$$

The number of 3-term progressions in A_1 is $\approx \alpha^3 N/4$, and so it follows that the number of 4-term progressions in A_1 is $\approx 2\alpha^3/27$.

The analysis of A_2 is rather more complicated. However one may check that if $|\{x\sqrt{3}\}|, |\{d\sqrt{3}\}| \leqslant 1/10$ and if $|\{y\sqrt{2}\{y\sqrt{3}\}\}| \leqslant \alpha/10$ for $y = x, x + d, x + 2d$, then $x + 3d \in A_2$. One can show that there are $\gg \alpha^3 N^2$ choices of x, d satisfying these constraints, and hence once again A_2 contains $\gg \alpha^3 N^2$ 4-term progressions.

7. Generalised quadratic obstructions

We saw in the last section that the set of linear exponentials $e(\theta n)$ is not a characteristic system for 4-term progressions. There we saw examples involving quadratics $n^2\theta$ and generalised quadratics $n\theta_1\{n\theta_2\}$, and these must clearly be addressed by any generalisation of Propositions 6.1 and 6.2 to 4-term APs. Somewhat remarkably, these quadratic and generalised quadratic examples are in a sense the only ones.

Proposition 7.1. *Suppose that $f_1, f_2, f_3, f_4 \colon [N] \to [-1, 1]$ are four functions and that*

$$|\mathbb{E}_{\substack{x_1,x_2,x_3,x_4 \\ x_1-2x_2+x_3=0 \\ x_2-2x_3+x_4=0}} f_1(x_1) f_2(x_2) f_3(x_3) f_4(x_4)| \geqslant \delta. \tag{16}$$

Then for any $i = 1, 2, 3, 4$ there is a generalised quadratic polynomial

$$\phi(n) = \sum_{r,s \leqslant C_1(\delta)} \beta_{rs}\{\theta_r n\}\{\theta_s n\} + \gamma_r\{\theta_r n\}, \tag{17}$$

where $\beta_{rs}, \gamma_r, \theta_r \in \mathbb{R}$, such that

$$|\mathbb{E}_{n \leqslant N} f_i(n) e(\phi(n))| \geqslant c_2(\delta).$$

We can take $C_1(\delta) \sim \exp(\delta^{-C})$ and $c_2(\delta) \sim \exp(-\delta^{-C})$.

Note that

$$\theta n^2 = 100\theta N^2 \left\{ \frac{n}{10N} \right\}^2$$

and

$$\theta_1 n\{\theta_2 n\} = 10\theta_1 N \left\{ \frac{n}{10N} \right\} \{\theta_2 n\}$$

for $n \leqslant N$, and so the phases which can be written in the form (17) do include all those which were discovered to be relevant in the preceding section.

The proof of Proposition 7.1 is given in [13]. It builds on earlier work of Gowers [8], [10]. In [13] (see also [14]) several results of a related nature are given, in which other characteristic systems for the equation $x_1 - 2x_2 + x_3 = x_2 - 2x_3 + x_4 = 0$ are given. These systems all have a "quadratic" flavour. We will discuss the family of 2-*step nilsequences*, which is perhaps the most conceptually appealing, in §9. In §11 we will mention the family of *local quadratics*, which are useful for computations involving the Möbius function. The only real merit of the generalised quadratic phases $e(\phi(n))$ discussed above is that they are easy to describe from first principles.

8. The Gowers norms and inverse theorems

The proof of Proposition 7.1 is long and complicated: there does not seem to be anything so simple as Formula (3) in the world of 4-term progressions. Very roughly

speaking one assumes that (16) holds, and then one proceeds to place more and more structure on each function f_i until eventually one establishes that f_i correlates with a generalised quadratic phase. There is a finite field setting for this argument, and we would recommend that the interested reader read this first: it may be found in [13, Ch. 5]. The ICM lecture of Gowers [9] is a fine introduction to the ideas in his paper [8], which is the foundation of our work.

There is only one part of the existing theory which we feel sure will play some rôle in future incarnations of these methods. This is the first step in the long series of deductions from (16), in which one shows that each f_i has large *Gowers norm*. For the purposes of this exposition[1] we define the Gowers U^2-norm $\|f\|_{U^2}$ of a function $f : [N] \to [-1, 1]$ by

$$\|f\|_{U^2}^4 := \mathbb{E}_{\substack{x_{00},x_{01},x_{10},x_{11}\leqslant N \\ x_{00}+x_{11}=x_{01}+x_{10}}} f(x_{00}) f(x_{01}) f(x_{10}) f(x_{11}),$$

which is a sort of average of f over two dimensional parallelograms. The U^k norm, $k \geqslant 3$, is an average of f over k-dimensional parallelepipeds. Written down formally it looks much more complicated than it is:

$$\|f\|_{U^k}^{2^k} := \mathbb{E}_{\substack{x_{0,\dots,0},\dots,x_{1,\dots,1} \\ x_{\omega(1)}+x_{\omega(2)}=x_{\omega(3)}+x_{\omega(4)}}} f(x_{0,\dots,0}) \dots f(x_{1,\dots,1}),$$

where there are 2^k variables x_ω, $\omega = (\omega_1, \dots, \omega_k) \in \{0, 1\}^k$, and the constraints range over all quadruples $(\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, \omega^{(4)}) \in (\{0, 1\}^k)^4$ with $\omega^{(1)}+\omega^{(2)} = \omega^{(3)}+\omega^{(4)}$.

The Gowers U^k norm governs the behaviour of any non-degenerate system $\boldsymbol{A}\boldsymbol{x} = 0$ in which $\boldsymbol{A}$ has $(k - 1)$ rows.

Proposition 8.1 (Generalised von Neumann theorem). *Suppose that $\boldsymbol{A}$ is a non-degenerate $s \times t$ matrix with integer entries. Suppose that $f_1, \dots, f_t : [N] \to [-1, 1]$ are functions and that*

$$|\mathbb{E}_{\substack{x_1,\dots,x_t \\ \boldsymbol{A}\boldsymbol{x}=0}} f_1(x_1) \dots f_t(x_t)| \geqslant \delta.$$

Then for each $i = 1, \dots, t$ we have

$$\|f_i\|_{U^{s+1}} \geqslant c_{\boldsymbol{A}}\delta.$$

The proof involves $s + 1$ applications of the Cauchy–Schwarz inequality. In this generality, the result was obtained in [14], though the proof technique is the same as in [10]. There are results in ergodic theory of the same general type, in which "non-conventional ergodic averages" are bounded using seminorms which are analogous to the U^k-norms: see [20].

[1] In practice we do all our work on the group $\mathbb{Z}/N'\mathbb{Z}$ for some prime $N' \geqslant N$ with $N' \approx M(\boldsymbol{A})N$, where $M(\boldsymbol{A})$ is some constant depending on the system of equations $\boldsymbol{A}\boldsymbol{x} = 0$ one is interested in. One advantage of this is that the number of solutions to $\boldsymbol{A}\boldsymbol{x} = 0$ in $\mathbb{Z}/N'\mathbb{Z}$ is much easier to count than the number of solutions in $[N]$. The Gowers norms defined here differ from the Gowers norms in those settings by constant factors, so for expository purposes they may be thought of as the same. In the group setting the constant $c_{\boldsymbol{A}}$ in Proposition 8.1 is simply 1.

Taking $s = k-2$ and $\boldsymbol{A}$ as in (1), we see that in particular the Gowers U^{k-1}-norm "controls" k-term progressions. The Gowers norms are, of course, themselves defined by a system of linear equations, and so they must be studied as part of a generalised Hardy–Littlewood method with as broad a scope as we would like. The Generalised von Neumann Theorem may be regarded as a statement to the effect that in a sense they represent the *only* systems of equations that need to be studied.

The Gowers norms do not feature in the classical Hardy–Littlewood method. It is, however, possible to prove a somewhat weaker version of Proposition 6.1 by combining the case $k = 3$ of Proposition 8.1 with the following *inverse theorem*:

Proposition 8.2 (Inverse theorem for U^2). *Suppose that N is large and that $f : [N] \to [-1, 1]$ is a function with $\|f\|_{U^2} \geqslant \delta$. Then we have*

$$\sup_{\theta\in[0,1)} |\mathbb{E}_{n\leqslant N} f(n)e(n\theta)| \geqslant 2\delta^2.$$

To prove this we note the formula

$$\mathbb{E}_{x_{00},x_{01},x_{10},x_{11}} f(x_{00})f(x_{01})f(x_{10})f(x_{11})1_{x_{00}+x_{11}=x_{01}+x_{10}} = \int_0^1 |\hat{f}(\theta)|^4\, d\theta,$$

where $\hat{f}(\theta) := \mathbb{E}_{n\leqslant N} f(n)e(n\theta)$. This implies that

$$\|f\|_{U^2}^4 = (3N + O(1))\|\hat{f}\|_4^4.$$

In view of the fact that $\|\hat{f}\|_2^2 \leqslant 1/N$, this and the assumption that $\|f\|_{U^2} \geqslant \delta$ imply that

$$\|\hat{f}\|_\infty^2 \geqslant (3 + o(1))\delta^4,$$

which implies the result.

This argument should be compared to the argument in (6), to which it corresponds rather closely.

To deduce Proposition 6.1 by passing through Proposition 8.2 is rather perverse, since the derivation is longer than the one that proceeds via an analogue of (3) and it leads to worse dependencies. With our current technology, however, this is the only method which is amenable to generalisation.

Similarly, one may deduce Proposition 7.1 from Proposition 8.1 and the following result.

Proposition 8.3 (Inverse theorem for the U^3-norm). *Suppose that $f : [N] \to \mathbb{R}$ is a function for which $\|f\|_\infty \leqslant 1$ and $\|f\|_{U^3} \geqslant \delta$. Then there is a generalised quadratic phase*

$$\phi(n) = \sum_{r,s\leqslant C_1(\delta)} \beta_{rs}\{\theta_r n\}\{\theta_s n\} + \gamma_r\{\theta_r n\}, \tag{18}$$

where $\beta_{rs}, \gamma_r, \theta_r \in \mathbb{R}$, such that

$$|\mathbb{E}_{n\leqslant N} f(n)e(\phi(n))| \geqslant c_2(\delta).$$

We can take $C_1(\delta) \sim \exp(\delta^{-C})$ and $c_2(\delta) \sim \exp(-\delta^{-C})$.

This result (and variations of it involving other "quadratic families") is in fact the main theorem in [13].

As we mentioned, one may find a series of seminorms which are analogous to the Gowers norms in the ergodic-theoretic work of Host and Kra [20]. There are no such seminorms in the related work of Ziegler [35], however, and this suggests that (as in the classical case) the Gowers norms may not be completely fundamental to a generalised Hardy–Littlewood method.

9. Nilsequences

In the previous section we introduced the Gowers U^k-norms, and stated inverse theorems for the U^2- and U^3-norms. These inverse theorems provide lists of rather algebraic functions which are *characteristic* for a given system of equations $\boldsymbol{A}\boldsymbol{x} = 0$. Roughly speaking, the linear phases $e(\theta n)$ are characteristic for single linear equations in which $\boldsymbol{A}$ is a $1 \times t$ matrix. Generalised quadratic phases $e(\phi(n))$ are characteristic for pairs of linear equations in which $\boldsymbol{A}$ is a non-degenerate $2 \times t$ matrix.

These two results leave open the question of whether there is a similar list of functions which is characteristic for the U^k-norm, $k \geqslant 4$ and hence, by the Generalised von Neumann Theorem, for non-degenerate systems defined by an $s \times t$ matrix with $s \geqslant 3$. The form of Propositions 8.2 and 8.3 does not suggest a particularly natural form for such a result, however, and indeed Proposition 8.2 is already rather unnatural-looking.

To make more natural statements, we introduce a class of functions called *nilsequences*.

Definition 9.1. Let G be a connected, simply connected, k-step nilpotent Lie group. That is, the central series $G_0 := G$, $G_{i+1} = [G, G_i]$ terminates with $G_k = \{e\}$. Let $\Gamma \subseteq G$ be a discrete, cocompact subgroup. The quotient G/Γ is then called a k-step nilmanifold. The group G acts on G/Γ via the map $T_g(x\Gamma) = xg\Gamma$. If $F\colon G/\Gamma \to \mathbb{C}$ is a bounded, Lipschitz function and $x \in G/\Gamma$ then we refer to the sequence $(F(T_g^n \cdot x))_{n\in\mathbb{N}}$ as a k-step nilsequence.

By analogy with the results of Host and Kra [20] in ergodic theory, we expect the collection of $(k-1)$-step nilsequences to be characteristic for the U^k-norm. The following conjecture is one of the guiding principles of the generalised Hardy–Littlewood method.

Conjecture 9.2 (Inverse conjecture for U^k-norms). Suppose that $k \geqslant 2$ and that $f\colon [N] \to [-1, 1]$ has $\|f\|_{U^k} \geqslant \delta$. Then there is a $(k-1)$-step nilmanifold G/Γ with dimension at most $C_{1,k}(\delta)$, together with a function $F\colon G/\Gamma \to \mathbb{C}$ with $\|F\|_\infty \leqslant 1$ and Lipschitz constant at most $C_{2,k}(\delta)$ and elements $g \in G$, $x \in G/\Gamma$ such that

$$|\mathbb{E}_{n\leqslant N} f(n)F(T_g^n \cdot x)| \geqslant c_{3,k}(\delta). \tag{19}$$

We can at least be sure that Conjecture 9.2 is no more complicated than necessary, since in [13, Ch. 12] we showed that if a bounded function f correlates with a $(k-1)$-step nilsequence as in (19) then f *does* have large Gowers U^k-norm. This, incidentally, is another reason to believe that the Gowers norms play a fundamental rôle in the theory. It is not the case that correlation of a function f with a $(k-1)$-step nilsequence prohibits f from enjoying cancellation along k-term arithmetic progressions, for example. In the case $k = 3$ an example of this phenomenon is given by the function f which equals α for $1 \leqslant n \leqslant N/3$ and -1 for $N/3 < n \leqslant N$, where α is the root between 1 and 2 of $\alpha^3 - \alpha^2 + 3\alpha - 4 = 0$. This f correlates with the constant nilsequence 1 yet exhibits cancellation along 3-term progressions, as the reader may care to check.

Conjecture 9.2 seems, at first sight, to be completely unrelated to Propositions 8.2 and 8.3. However after a moment's thought one realises that a linear phase $e(\theta n)$ can be regarded as a 1-step nilsequence in which $G = \mathbb{R}$, $\Gamma = \mathbb{Z}$, $g = \theta$ and $x = 0$. Thus Proposition 8.2 immediately implies the case $k = 2$ of Conjecture 9.2.

The case $k = 3$ is proved in [13]. One first proves Proposition 8.3, and then one shows how any generalised quadratic phase $e(\phi(n))$ may be approximated by a 2-step nilsequence. Let us discuss a simple example, the *Heisenberg nilmanifold*, to convince the reader that 2-step nilsequences can give rise to "generalised quadratic" behaviour.

Example 9.1 (The Heisenberg nilmanifold). Consider

$$G := \begin{pmatrix} 1 & \mathbb{R} & \mathbb{R} \\ 0 & 1 & \mathbb{R} \\ 0 & 0 & 1 \end{pmatrix}; \quad \Gamma := \begin{pmatrix} 1 & \mathbb{Z} & \mathbb{Z} \\ 0 & 1 & \mathbb{Z} \\ 0 & 0 & 1 \end{pmatrix}.$$

Then G/Γ is a 2-step nilmanifold. By using the identification

$$(x, y, z) \equiv \begin{pmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{pmatrix} \Gamma,$$

we can identify G/Γ (as a set) with $\mathbb{R}^3$, quotiented out by the equivalence relations

$$(x, y, z) \sim (x + a, y + b + cx, z + c) \quad \text{for all } a, b, c \in \mathbb{Z}.$$

This can in turn be coordinatised by the cylinder $(\mathbb{R}/\mathbb{Z})^2 \times [-1/2, 1/2]$ with the identification $(x, y, -1/2) \sim (x, x + y, 1/2)$. Let $F\colon G/\Gamma \to \mathbb{C}$ be a function. We may lift this to a function $\widetilde{F}\colon G \to \mathbb{C}$, defined by $\widetilde{F}(g) := F(g\Gamma)$. In coordinates, this lift takes the form

$$\widetilde{F}(x, y, z) = F(x \pmod 1, y - [z]x \pmod 1, \{z\})$$

where $[z] = z - \{z\}$ is the nearest integer to x. Let

$$g := \begin{pmatrix} 1 & \alpha & \beta \\ 0 & 1 & \gamma \\ 0 & 0 & 1 \end{pmatrix}$$

be an element of G. Then the shift $T_g : G \to G$ is given by

$$T_g(x, y, z) = (x + \alpha, y + \beta + \gamma x, z + \gamma).$$

A short induction confirms, for example, that

$$T_g^n(0, 0, 0) = (n\alpha, n\beta + \tfrac{1}{2}n(n+1)\alpha\gamma, n\gamma).$$

Therefore if $F : G/\Gamma \to G/\Gamma$ is any Lipschitz function, written as a function $F : (\mathbb{R}/\mathbb{Z})^2 \times [-1/2, 1/2] \to \mathbb{C}$ with $F(-1/2, y, z) = F(1/2, y + z, z)$, then we have

$$F(T_g^n(0, 0, 0)) = F(n\alpha \pmod 1, n\beta + \tfrac{1}{2}n(n+1)\alpha\gamma - [n\gamma]n\alpha \pmod 1, \{n\gamma\}).$$

The term $[n\gamma]n\alpha$ which appears here certainly exhibits a sort of generalised quadratic behaviour. For a complete description of how an arbitrary generalised quadratic phase $e(\phi(n))$ can be approximated by a two-step nilsequence, we refer the reader to [13, Ch. 12].

Let us conclude this section by stating, for the reader's convenience, a result/conjecture which summarises much of our discussion so far in one place.

Theorem 9.3 (Green–Tao [13]). *We have the following two statements.*

(i) (Generalised von Neumann) *Suppose that s and t are positive integers with $s + 2 \leqslant t$. Suppose that $\boldsymbol{A}$ is a non-degenerate $s \times t$ matrix with integer entries. Suppose that $f_1, \dots, f_t : [N] \to [-1, 1]$ are functions and that*

$$|\mathbb{E}_{\substack{x_1,\dots,x_t \\ \boldsymbol{A}\boldsymbol{x}=0}} f_1(x_1) \dots f_t(x_t)| \geqslant \delta. \tag{20}$$

Then for each $i = 1, \dots, t$ we have

$$\|f_i\|_{U^{s+1}} \geqslant c_{\boldsymbol{A}}\delta.$$

(ii) (Gowers inverse result: proved for $k = 2, 3$, conjectural for $k \geqslant 4$) *Suppose that $f : [N] \to [-1, 1]$ has $\|f\|_{U^k} \geqslant \delta$. Then there is a $(k-1)$-step nilmanifold G/Γ with dimension at most $C_{1,k}(\delta)$, together with a function $F : G/\Gamma \to \mathbb{C}$ with $\|F\|_\infty \leqslant 1$ and Lipschitz constant at most $C_{2,k}(\delta)$ and elements $g \in G$, $x \in G/\Gamma$ such that*

$$|\mathbb{E}_{n \leqslant N} f(n) F(T_g^n \cdot x)| \geqslant c_{3,k}(\delta). \tag{21}$$

In particular when $s = 1$ or 2 and (20) holds for some $\boldsymbol{A}$ and some δ then for each $i = 1, \dots, t$ there is a 2-step nilsequence $(F(T_g^n \cdot x))_{n \in \mathbb{N}}$ such that

$$|\mathbb{E}_{n \leqslant N} f_i(n) F(T_g^n \cdot x)| \geqslant c_{\boldsymbol{A}}(\delta). \tag{22}$$

10. Working with the primes

Let us suppose that we wish to count four-term progressions in the primes. One might try to apply Theorem 9.3 with the functions f_i equal to the balanced function of A, the set of primes $p \leqslant N$, and then hope to rule out a correlation such as (21) for some $\delta = o(\alpha^t)$ (here, of course, $\alpha \approx \log^{-1} N$ by the prime number theorem). This would then lead to an asymptotic using various instances of (20) together with the triangle inequality.

There are two reasons why this is a hopeless strategy. First of all, the primes *do* correlate with nilsequences. In fact since all primes other than 2 are odd it is easy to see that

$$\mathbb{E}_{n \leqslant N} f_A(n) e(n/2) \approx -\alpha.$$

There is a way to circumvent this problem, which we call the W-trick. The idea is that if $W = 2 \times 3 \times \cdots \times w(N)$ is the product of the first several primes, then for any b coprime to W the set

$$A_b := \{n \leqslant N : Wn + b \text{ is prime}\}$$

does not exhibit significant bias in progressions with common difference $q \leqslant w(N)$. One can then count 4-term progressions in the primes by counting 4-term progressions in $A_{b_1} \times \ldots A_{b_4}$ for each quadruple $(b_1, \ldots, b_4) \in (\mathbb{Z}/W\mathbb{Z})^{\times 4}$ in arithmetic progression and adding.

We refer to any set A_b as a set of "W-tricked primes". In practice one is only free to take $w(N) \sim \log\log N$, since one must be able to understand the distribution of primes in progressions with common difference W (note that even on GRH one could only take $w(N) \sim c \log N$). Even assuming we could obtain optimal results concerning the correlation of the W-tricked primes with 2-step nilsequences, this information will be very weak indeed.

This highlights a more serious problem with the suggested strategy. Suppose that $A \subseteq [N]$ is a set of density α for which there is no obvious reason why A should have an unexpectedly large or small number of 4-term APs, that is to say for which we might hope to prove that

$$\mathbb{E}_{\substack{x_1,x_2,x_3,x_4 \\ x_1-2x_2+x_3=0 \\ x_2-2x_3+x_4=0}} 1_A(x_1)1_A(x_2)1_A(x_3)1_A(x_4) \approx \alpha^4. \tag{23}$$

For example, A might be the W-tricked primes less than N, in which case $\alpha \sim \frac{W}{\phi(W)} \log^{-1} N$.

We might prove (23) by writing $1_A = \alpha + f_A$, expanding as the sum of sixteen terms, and showing that fifteen of these are $o(\alpha^4)$ by appealing to Theorem 9.3, and ruling out a correlation with a 2-step nilsequence as in (22). Unfortunately we will be operating with $\delta = o(\alpha^4) \ll \log^{-4+\varepsilon} N$, and the dependence of $c_A(\delta)$ on δ is very weak, being of the form $\exp(-\delta^{-C})$. Thus we are asking to rule out the possibility

that

$$|\mathbb{E}_{n\leqslant N} f_A(n) F(T_g^n \cdot x)| \gg \exp(-\log^C N)$$

for some potentially rather large C. This is a problem, since one would never expect more than square root cancellation in any such expression. In fact for the W-tricked primes one only has a small amount (depending on $w(N)$) of potential cancellation to work with and to all intents and purposes one should not bank on having available any estimate stronger than

$$\mathbb{E}_{n\leqslant N} f_A(n) F(T_g^n \cdot x) = o(1).$$

What one really needs is a version of Proposition 16 which applies to functions which need not be bounded by 1. Then one could hope to work with the von Mangoldt function Λ instead of the far less natural characteristic function 1_A, or more accurately with W-tricked variants of the von Mangoldt function such as

$$\Lambda_{b,W}(n) := \frac{\phi(W)}{W}\Lambda(Wn + b).$$

Such a result is the main result of our forthcoming paper [15]. It would take us too far afield to say anything concerning its proof, other than that it uses one of the key tools from our paper [12] on long progressions of primes, the "ergodic transference" technology of [12, Chs. 6,7,8].

Proposition 10.1 (Transference principle, [15]). *Suppose that* $\nu\colon [N] \to \mathbb{R}^+$ *is a* pseudorandom measure. *Then*

(i) *The generalised von Neumann theorem, Theorem* 9.3 (i)*, continues to hold for functions* $f_1, \dots, f_t\colon [N] \to \mathbb{R}^+$ *such that* $|f_i(x)| \leqslant 1 + \nu(x)$ *pointwise* (*the value of* $c_{\boldsymbol{A}}$ *may need to be reduced slightly*).

(ii) *If the Gowers inverse conjecture, Theorem* 9.3 (ii)*, holds for a given value of* k *then it continues to hold for a function* f *such that* $|f(x)| \leqslant 1+\nu(x)$ *pointwise. In particular such an extension of the Gowers inverse conjecture is true when* $k = 2, 3$.

The reader may consult [12, Ch. 3] for a definition of the term *pseudorandom measure* and a discussion concerning it. For the purposes of this article the reader can merely accept that there is such a notion, and furthermore that one may construct a pseudorandom measure $\nu\colon [N] \to \mathbb{R}^+$ such that $\nu+1$ dominates any fixed W-tricked von Mangoldt function $\Lambda_{W,b}$. The construction of ν comes from sieve theoretic ideas originating with Selberg. The confirmation that ν is pseudorandom is essentially due, in a very different context, to Goldston and Yıldırım [7].

Applying these two results, one may see that Conjecture 2.1 for a given non-degenerate $s \times t$ matrix $\boldsymbol{A}$ is a consequence of the Gowers inverse conjecture in the case $k = s + 1$ together with a bound of the form

$$\mathbb{E}_{n\leqslant N}(\Lambda_{b,W}(n) - 1) F(T_g^n \cdot x) = o_{G/\Gamma,F}(1) \tag{24}$$

for every s-step nilsequence $(F(T_g^n \cdot x))_{n\in\mathbb{N}}$.

By effecting a decomposition of $\Lambda_{b,W}$ as $\Lambda_{b,W}^\sharp + \Lambda_{b,W}^\flat$ rather like that in §4, the proof of this statement may be further reduced to a similar result for the Möbius function:

Conjecture 10.2 (Möbius and nilsequences). For all $A > 0$ we have the bound

$$\mathbb{E}_{n\leqslant N}\mu(n)F(T_g^n \cdot x) \ll_{A,G/\Gamma,F} \log^{-A} N$$

for every k-step nilsequence $(F(T_g^n \cdot x))_{n\in\mathbb{N}}$.

Note that we require more cancellation (a power of a logarithm) here than in (24). This is because in passing from μ to $\Lambda_{b,W}^\flat$ one loses a logarithm in performing partial summation as in the derivation of (7). The method we have in mind to prove Conjecture 10.2, however, is likely to give this strong cancellation at no extra cost.

Conjecture 10.2 posits a rather vast generalisation of Davenport's bound. The conjecture is, of course, highly plausible in view of the Möbius randomness law.

Let us remark that the derivation of (24) from Conjecture 10.2 is not at all immediate, since one must also handle the contribution from $\Lambda_{b,W}^\sharp$. To do this one uses methods of classical analytic number theory rather similar to those of Goldston and Yıldırım [7].

11. Möbius and nilsequences

The main result of [14] is a proof of Conjecture 10.2 in the case $k = 2$. This leads, by the reasoning outlined in the previous section, to a proof of Conjecture 2.4 in the case $s = 2$.

We remarked that the classical Hardy–Littlewood method was a technique of harmonic analysis. We also highlighted the idea of dividing into major and minor arcs. We have said much on the subject of generalising the underlying harmonic analysis, but as yet there has been nothing said about a suitable extension of major and minor arcs. In this section we describe such an extension by making some remarks concerning the proof of the case $k = 2$ of Conjecture 10.2.

In §5 we discussed how bounds on Type I and II sums may be used to show that a given function f does not correlate with Möbius. Recalling our "inverse" strategy for proving Davenport's bound, one might be tempted to go straight into Proposition 5.1 with $f(n) = F(T_g^n \cdot x)$, a 2-step nilsequence, posit largeness of either a Type I or a Type II sum, and then use this to say that the nilsequence is somehow "major arc". One might then hope to handle the major nilsequences by some other method, perhaps the theory of L-functions.

Such an attempt is a little too simplistic, for the following reason. Returning to the 1-step case, note that the sum of two 1-step nilsequences is also a 1-step nilsequence (on the product nilmanifold $G_1/\Gamma_1 \times G_2/\Gamma_2$). In particular, the function

$f(n) = e(n/5) + e(n\sqrt{2})$ is a 1-step nilsequence. We know, however, that to handle correlation of Möbius with $e(n/5)$ we need to know something about L-functions, whereas we do not have an L-function method of handling $e(n\sqrt{2})$. This suggests that some sort of preliminary decomposition of the function f is in order, and such a suggestion turns out to be correct.

In the 2-step case, a nilsequence $F(T_g^n \cdot x)$ can be decomposed into *local quadratics*. These are objects of the form

$$f(n) := 1_{B_N}(n)e(\phi(n)), \tag{25}$$

where B_N is a set of the form

$$B_N := \{n : N/2 \leqslant n < N : F_1(n) \neq 0\}$$

for some 1-step nilsequence F_1 depending on F, G/Γ, g and x, and $\phi\colon B_N \to \mathbb{R}/\mathbb{Z}$ is *locally quadratic*. This means that one may unambiguously define the second derivative $\phi''(h_1, h_2)$ to equal

$$\phi(x + h_1 + h_2) - \phi(x + h_1) - \phi(x + h_2) + \phi(x)$$

for any x such that $x, x + h_1, x + h_2, x + h_1 + h_2 \in B_N$.

It turns out that for the purposes of analysing Type I and II sums the cutoff 1_{B_N} plays a subservient rôle. The phase ϕ, on the other hand, is crucial. The bulk of [14] is devoted to showing that if either a Type I or a Type II sum involving some f as in (25) is large, then ϕ is *major arc*. This is a direct analogue of the proof of Davenport's bound as phrased at the end of §5 (the "inverse" approach). Roughly speaking, ϕ is said to be major arc if $q\phi''(h_1, h_2)$ is small for some smallish q and all h_1, h_2, which in turn essentially means that ϕ is slowly varying on B_N intersected with any fixed progression $a \pmod q$. For a detailed discussion see [14]. Suffice it to say that the passage from large Type I/II sum to ϕ being major arc is long and difficult, and requires many applications of the Cauchy–Schwarz inequality to manipulate the phase ϕ into a helpful form, as well as basic tools of equidistribution such as a version of the Erdős–Turán inequality.

Recalling Proposition 5.1, one has reduced the case $s = 2$ of Conjecture 10.2 to the statement that

$$\mathbb{E}_{n \leqslant N} \mu(n) 1_{B_N}(n) e(\phi(n)) \ll_A \log^{-A} N$$

for any major arc phase ϕ. It turns out that $1_{B_N}(n)e(\phi(n))$ can, in this case, be closely approximated by a sum of linear phases $e(\theta n)$, and so we may conclude using Proposition 4.1.

Note that this analysis has the flavour of an induction on s, the step of the nilsequence we are considering. We expect to see this more clearly when addressing the general case of Conjecture 10.2 in future work.

12. Future directions

The most obvious avenue of research left open is to generalise everything we have done for $s = 2$ to the case $s \geqslant 3$. In particular we would like inverse theorems for the U^k-norms for $k \geqslant 4$, and a proof of Conjecture 10.2 for $s \geqslant 3$. We are currently working towards this goal. We expect that the methods of Gowers [10] can be adapted to achieve the inverse theorem, though this will not be straightforward. It is also very likely that the "inverse" approach to handling Type I and II sums can be adapted to the higher-step case of Conjecture 10.2, though again we do not expect this to be wholly straightforward.

It would be very desirable to have good bounds for error terms such as the $o(1)$ in Theorem 2.5. We are sure that our current estimate for the error in Theorem 2.5 is the worst that has ever featured in analytic number theory – the error term is a *completely ineffective* $o(1)$! Ultimately this is because to show that the error is less than δ one finds oneself needing to rule out a real zero of some $L(s, \chi)$, χ a primitive quadratic character to the modulus q, with $s > 1 - Cq^{-\varepsilon}$, where $\varepsilon = \varepsilon(\delta) \to 0$ as $\delta \to 0$. Siegel's theorem states that for any $\varepsilon > 0$ there is such a C, but it is, of course, not possible to specify C effectively.

It is clear that the spectre of ineffectivity does not rear its head under the assumption of GRH, and we believe that our methods lead to an error term of the form $\log^{-c} N$ in Theorem 2.5.

There are other, presumably more tractible, ways in which one might obtain an explicit error term. Improvements to the combinatorial tools used in [13], particularly advances on the circle of conjectures known as the "polynomial Freiman–Ruzsa conjecture", could be very helpful here.

We turn now to goals which lie further away. I have hinted at various places in this survey that the way in which we see nilsequences arising is very long-winded and, presumably, not the "right" way. The ergodic theorists [20], [35] do admittedly discover the rôle of these functions somewhat less painfully (albeit after setting up a good deal of notation). Nilsequences seem such natural objects, however, that there ought to be a much better way of appreciating their place in the study of systems of linear equations. Recalling that $\|f\|_{U^2}$ is essentially the L^4 norm of $\hat{f}$ one might even ask, for example,

Question 12.1. Is there a usable "formula" relating $\|f\|_{U^3}$ and certain of the "nil-fourier coefficients" $\mathbb{E}_{n \leqslant N} f(n) F(T_g^n \cdot x)$?

Such a formula would assuredly have to be very exotic on account of the vast profusion of nilsequences which might enter into consideration. The nilsequences are not naturally parametrised by anything so simple as the circle S^1, which gave its name to the classical circle method.

Let us conclude with some speculations on non-linear systems of equations, where our knowledge is at present essentially non-existent. We have seen in Conjecture 9.2 that the behaviour of an any system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is non-degenerate in the sense

of Definition 2.3, should be governed by a very "hard" or "algebraic" collection of *characteristic functions,* in this case the nilsequences.

On the other hand degenerate linear systems, such as $x_1 - x_2 = 1$, do not have this property. To see this, suppose that $N = 2m$ is even and let $A \subseteq [N]$ be a set formed by setting $A \cap \{2i, 2i+1\} = \{2i\}$ or $\{2i+1\}$, these choices being independent in i for $i = 0, \ldots, m-1$. Then $|A| = N/2$, and A is indistinguishable from a truly random set by taking inner products with any conceivable "hard" character such as a linear or quadratic phase. However, A is expected to have about $N/8$ solutions to $x_1 - x_2 = 1$, whereas a random set has about twice this many.

One might call an equation or system of equations for which a "hard" characteristic system exists a *mixing* system. We do not have a precise definition of this notion. Some non-linear equations are known to be mixing – for example, the linear phases $e(\theta n)$ form a characteristic system for the equation $x_1 + x_2 = x_3^2$. Many more are not. It would be very interesting to know, for example, whether the equation $x_1x_2 - x_3x_4 = 1$ is mixing and, if so, what a characteristic system for it might be. This seems to be a very difficult question as the analysis of this equation even in very specific situations involves deep methods from the theory of automorphic forms.

References

[1] Baker, R. C. and Harman, G., Exponential sums formed with the Möbius function. *J. London Math. Soc.* (2) **43** (2) (1991), 193–198.

[2] Balog, A., Linear equations in primes. *Mathematika* **39** (1992), 367–378.

[3] Bourgain, J., On triples in arithmetic progression. *Geom. Funct. Anal.* **9** (1999), 968–984.

[4] Chowla, S., There exists an infinity of 3-combinations of primes in A.P. *Proc. Lahore. Philos. Soc.* **6** (2) (1944), 15–16.

[5] Davenport, H., *Multiplicative number theory*. Third edition, Grad. Texts in Math. 74, Springer-Verlag, New York 2000.

[6] Furstenberg, H. and Weiss, B., A mean ergodic theorem for $1/N \sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. In *Convergence in ergodic theory and probability* (Columbus, OH, 1993), Ohio State Univ. Math. Res. Inst. Publ. 5, Walter de Gruyter, Berlin 1996, 193–227.

[7] Goldston, D. A. and Yıldırım, C. Y., Small gaps between primes, I. Preprint.

[8] Gowers, W. T., A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geom. Funct. Anal.* **8** (1998), 529–551.

[9] Gowers, W. T., Fourier analysis and Szemerédi's theorem. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. I, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 617–629.

[10] Gowers, W. T., A new proof of Szemerédi's theorem. *Geom. Funct. Anal.* **11** (2001), 465–588.

[11] Green, B. J., Long arithmetic progressions of primes. Preprint, submitted to *Proceedings of the Gauss-Dirichlet Conference*, Göttingen 2005.

[12] Green, B. J. and Tao, T. C., The primes contain arbitrarily long arithmetic progressions. *Ann. of Math.*, to appear.

[13] Green, B. J. and Tao, T. C., An inverse theorem for the Gowers U^3-norm, with applications. Submitted.

[14] Green, B. J. and Tao, T. C., Quadratic uniformity of the Möbius function. Preprint.

[15] Green, B. J. and Tao, T. C., Linear equations in primes. In preparation.

[16] Hardy, G. H. and Littlewood, J. E., Some problems of "Partitio Numerorum". III. On the expression of a number as a sum of primes. *Acta. Math.* **44** (1923), 1–70.

[17] Hardy, G. H. and Littlewood, J. E., Some problems of "Partitio Numerorum". V. A further contribution to the study of Goldbach's problem. *Proc. London Math. Soc.* (2) **22** (1923), 46–56.

[18] Hardy, G. H. and Ramanujan, S., Asymptotic formulæ in combinatory analysis. *Proc. London Math. Soc.* (2) **17** (1918), 75–115.

[19] Heath-Brown, D. R. Three primes and an almost prime in arithmetic progression. *J. London Math. Soc.* (2) **23** (1981), 396–414.

[20] Host, B. and Kra, B. Non-conventional ergodic averages and nilmanifolds. *Ann. of Math.* **161** (1) (2005), 397–488.

[21] Iwaniec, H. and Kowalski, E. *Analytic number theory.* Amer. Math. Soc. Colloq. Publ. 53, Amer. Math. Soc., Providence, RI, 2004.

[22] Iwaniec, H., Luo, W and Sarnak, P., Low lying zeroes of families of L-functions. *Inst. Hautes Études Sci. Publ. Math.* **91** (2000), 55–131.

[23] Kra, B., The Green-Tao Theorem on arithmetic progressions in the primes: an ergodic point of view. *Bull. Amer. Math. Soc.* **43** (2006), 3–23.

[24] Kra, B., From combinatorics to ergodic theory and back again. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume III, EMS Publishing House, Zürich 2006, 57–76.

[25] Kumchev, A. V. and Tolev, D. I., An invitation to additive prime number theory. *Serdica Math. J.* **31** (1–2) (2005), 1–74.

[26] Salem, R. and Zygmund, A., Some properties of trigonometric series whose terms have random signs. *Acta Math.* **91** (1954), 245–301.

[27] Tao, T. C., Arithmetic progressions and the primes. *El Escorial Proceedings* 2004, to appear.

[28] Tao, T. C., Obstructions to uniformity, and arithmetic patterns in the primes. Preprint.

[29] Van der Corput, J. G., Über Summen von Primzahlen und Primzahlquadraten. *Math. Ann.* **116** (1939), 1–50.

[30] Vaughan, R. C., Sommes trigonométriques sur les nombres premiers, *C. R. Acad. Sci. Paris Sér. A* **285** (16) (1977), A981–A983.

[31] Vaughan, R. C., Hardy's Legacy to Number Theory. *J. Austral. Math. Soc. Ser. A* **65** (1998), 238–266.

[32] Vaughan, R. C. and Wooley, T. D., Waring's problem: a survey. In *Number theory for the millennium, III* (Urbana, IL, 2000), A K Peters, Natick, MA, 2002, 301–340.

[33] Vinogradov, I. M., Representation of an odd number as the sum of three primes. *Dokl. Akad. Nauk SSSR* **15** (1937), 291–294.

[34] Wooley, T. D., Diophantine problems in many variables: the rôle of additive number theory. In *Topics in Number Theory* (ed. by S. D. Ahlgren et al.), Kluwer Academic Publishers, Dordrecht 1999, 49–83.

[35] Ziegler, T., Universal characteristic factors and Furstenberg averages. *J. Amer. Math. Soc.*, to appear.

School of Mathematics, University Walk, Bristol BS8 1TW, England
E-mail: b.j.green@bristol.ac.uk

Aspects géométriques du Lemme Fondamental de Langlands-Shelstad

Gérard Laumon*

Abstract. In order to compute the Hasse–Weil zeta functions of the Shimura varieties or to establish some cases of the Langlands functoriality one first needs to stabilize the Arthur–Selberg trace formula. This stabilization can be done only if some combinatorial identities between orbital integrals over p-adic reductive groups are satisfied. The series of these conjectural identities form the so-called "Fundamental Lemma". We present here some key points of the geometric approaches which have been used by Goresky, Kottwitz and MacPherson on the one hand and Ngô Bao Châu and myself one the other hand, to prove some cases of the Fundamental Lemma.

Résumé. Pour calculer les fonctions zêta de Hasse-Weil des variétés de Shimura ou pour établir certains cas de la fonctorialité de Langlands, il faut dans un premier temps stabiliser la formule des traces d'Arthur-Selberg. Cette stabilisation n'est possible que si certaines identités combinatoires entre intégrales orbitales sur les groupes réductifs p-adiques sont vérifiées. Ces identités conjecturales ont été regroupées sous la terminologie générique de "Lemme Fondamental". Nous présentons ici quelques points clé des approches géométriques utilisées par Goresky, Kottwitz et MacPherson d'une part, et Ngô Bao Châu et moi-même d'autre part, pour démontrer certains cas du Lemme Fondamental.

Mathematics Subject Classification (2000). Primary 14H60, 11F72 ; Secondary 22E35.

Keywords. Fundamental Lemma, endoscopy, Langlands functoriality.

Mots-clés. Lemme Fondamental, endoscopie, fonctorialité de Langlands.

1. Introduction

Le programme de Langlands est un faisceau de relations entre les représentations automorphes des groupes réductifs sur un corps local ou global et les représentations galoisiennes de ce même corps local ou global. Ce programme est très vaste puisqu'il contient une bonne partie de la théorie des nombres et aussi de la géométrie algébrique du fait de ses relations avec la théorie des motifs de Grothendieck, et il est donc très loin d'être achevé.

Nous nous intéressons ici à un aspect technique de ce programme. Pour calculer les fonctions zêta de Hasse-Weil des variétés de Shimura ou pour établir certains cas

*UMR 8628 du CNRS

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

de la fonctorialité de Langlands, il faut dans un premier temps stabiliser la formule des traces d'Arthur-Selberg (cf. [19], [12], [13] et [3]). Cette stabilisation n'est possible que si certaines identités combinatoires entre intégrales orbitales sur les groupes réductifs p-adiques sont vérifiées. Ces identités conjecturales ont été regroupées sous la terminologie générique et un peu baroque de "Lemme Fondamental".

Le Lemme Fondamental a été découvert par Labesse et Langlands dans leur travail sur la stabilisation de la formule des traces de Selberg pour SL(2) (cf. [18]). Langlands et Shelstad en ont donné une formulation générale toujours conjecturale dans [20]. Waldspurger en a ensuite formulé une variante pour les algèbres de Lie (cf. [28] et [16]) et Kottwitz et Shelstad ont donné une conjecture précise dans le cas "tordu" (cf. [17]). Pour être complet, signalons aussi les variantes "pondérée" et "pondérée tordue" du Lemme Fondamental conjecturées par Arthur (cf. [3]).

Dans certain cas il est possible de prouver le Lemme Fondamental en n'utilisant que des techniques combinatoires ou d'analyse harmonique. Cependant le cas général semble hors d'atteinte de cette manière et un recours à des techniques de la géométrie algébrique est probablement nécessaire.

Dans cet exposé nous présentons quelques points clé des approches géométriques utilisées par Goresky, Kottwitz et MacPherson d'une part, et Ngô Bao Châu et moi-même d'autre part, pour démontrer certains cas du Lemme Fondamental. Nous renvoyons aux articles originaux ([7] et [22]), à l'exposé de Ngô dans ce volume (cf. [23]) et à l'exposé Bourbaki de Dat (cf. [5]) pour plus de détails

2. Groupes réductifs

Soit F un corps local non archimédien, i.e. une extension finie de $\mathbb{Q}_p$ ou $\mathbb{F}_p((t))$. On note $\mathcal{O}_F$ l'anneau des entiers de F (la clôture intégrale de $\mathbb{Z}_p$ ou $\mathbb{F}_p[[t]]$ dans F) et $\mathfrak{p}_F$ l'idéal maximal de $\mathcal{O}_F$. Soit $\overline{F}$ une clôture algébrique de F.

Soit G un groupe (algébrique) réductif connexe sur F. On suppose que G est *quasi-déployé*, c'est-à-dire que G admet un sous-groupe de Borel défini sur F et qu'il se déploie au-dessus d'une extension finie non ramifiée $F' \subset \overline{F}$ de F. Notre groupe G est donc *non ramifié* et il existe des schémas en groupes lisses $\mathcal{G}$ sur le trait $\mathrm{Spec}(\mathcal{O}_F)$ dont la fibre générique est G et la fibre spéciale est un groupe réductif connexe sur le corps résiduel $\mathcal{O}_F/\mathfrak{p}_F$ de F. Soient $\mathcal{G}$ un tel schéma en groupes et $K = \mathcal{G}(\mathcal{O}_F) \subset G(F)$ le sous-groupe compact maximal *hyperspécial* correspondant.

Exemple 2.1. On suppose que la caractéristique résiduelle p de F est > 2. Soit F' l'extension quadratique non ramifiée de F contenue dans $\overline{F}$. On note τ l'élément non trivial du groupe de Galois $\mathrm{Gal}(F'/F)$.

On rappelle que le groupe $F^\times/\mathrm{Nr}_{F'/F}(F'^\times)$ est le groupe à deux éléments engendré par la classe de n'importe quelle uniformisante $\varpi_F \in \mathfrak{p}_F$ de F. On l'identifie à $\mathbb{Z}/2\mathbb{Z}$ dans la suite.

Un F'-espace hermitien est un espace vectoriel V de dimension finie sur F' muni d'une forme τ-sesquilinéaire symétrique non dégénérée Φ. Un tel F'-espace vectoriel hermitien (V, Φ) admet un discriminant : la classe dans $F^\times/\mathrm{Nr}_{F'/F}(F'^\times)$ du déterminant de la matrice de Φ dans une base de V. Deux F'-espaces hermitiens de même dimension sont isomorphes si et seulement s'ils ont même discriminant.

Le groupe unitaire des automorphismes

$$\mathrm{U}(V, \Phi) \subset \mathrm{Res}_{F'/F}\mathrm{Aut}_{F'}(V).$$

d'un tel F'-espace hermitien (V, Φ) est un groupe réductif connexe sur F. Il se déploie sur F' et est en fait une forme (extérieure) de $\mathrm{Aut}_{F'}(V)$ sur F. Il est quasi-déployé, et donc non ramifié, si et seulement si le discriminant de (V, Φ) est trivial. Dans ce cas, il existe des $\mathcal{O}_{F'}$-réseaux $M \subset V$ qui sont auto-duaux pour Φ au sens où le $\mathcal{O}_{F'}$-réseau

$$M^\perp = \{v \in V \mid \Phi(v, M) \subset \mathcal{O}_{F'}\} \subset V$$

est égal à M ; les sous-groupes maximaux hyperspéciaux sont alors les stabilisateurs dans $\mathrm{U}(V, \Phi)(F)$ de ces réseaux.

Soient (V, Φ) un F'-espace hermitien (V, Φ) de dimension n dont le discriminant est trivial et M un $\mathcal{O}_{F'}$-réseau auto-dual de V. On peut trouver une base de M sur $\mathcal{O}_{F'}$, qui est aussi une base de V sur F', dans laquelle la matrice de Φ est égale à la matrice Φ_n qui n'a pour seules entrées non nulles que les $(\Phi_n)_{i,n+1-i} = 1$ pour $i = 1, \ldots, n$. Tout couple formé d'un groupe unitaire non ramifié et d'une sous-groupe compact maximal hyperspécial est donc isomorphe à un groupe unitaire standard

$$\mathrm{U}(n) = \{g \in \mathrm{GL}_{F'}(n) \mid {}^t g^\tau \Phi_n g = \Phi_n\}$$

avec son sous-groupe compact maximal hyperspécial standard

$$\mathrm{U}(n)(F) \cap \mathrm{GL}(n, \mathcal{O}_{F'}).$$

Le groupe réductif $\mathrm{U}(n)$ ainsi défini admet pour groupe de Borel sur F le sous-groupe formé des $g \in \mathrm{U}(n)$ qui sont des matrices triangulaires supérieures dans $\mathrm{GL}_{F'}(n)$.

3. Intégrales orbitales

Soit $\mathfrak{g}$ l'algèbre de Lie de G qui est en particulier un espace vectoriel de dimension finie sur F. Notons $\mathrm{ad}\colon G \to \mathrm{Aut}_F(\mathfrak{g})$ la représentation adjointe de G. La classe de conjugaison d'un élément γ de $\mathfrak{g}$ est l'ensemble des $\mathrm{ad}(g)(\gamma)$ pour g parcourant $G(F)$; le centralisateur de γ est le sous-schéma en groupes $Z_G(\gamma)$ de G défini par la condition $\mathrm{ad}(g)(\gamma) = \gamma$. On rappelle que le centralisateur d'un élément régulier semi-simple de $\mathfrak{g}$ est toujours connexe.

Soit T un tore maximal de G que l'on suppose *anisotrope*, c'est-à-dire ne contenant pas de sous-tore déployé sur F non trivial, de sorte que le groupe $T(F)$ est compact. Soit $\mathfrak{t} \subset \mathfrak{g}$ l'algèbre de Lie de T.

Pour tout élément γ de $\mathfrak{t} \subset \mathfrak{g}$ qui est régulier, c'est-à-dire dont le centralisateur dans G est égal à T, toute fonction $f : \mathfrak{g} \to \mathbb{C}$ localement constante et à support compact et toute mesure de Haar $\mathrm{d}g$ sur $G(F)$, on peut former l'*intégrale orbitale* en γ de f

$$\mathrm{O}_\gamma^G(f, \mathrm{d}g) = \int_{G(F)} f(\mathrm{ad}(g^{-1})(\gamma))\,\mathrm{d}g.$$

On vérifie que l'ensemble

$$\{g \in G(F) \mid \mathrm{ad}(g^{-1})(\gamma) \in \mathrm{Supp}(f)\}$$

est compact et donc que cette intégrale orbitale converge (c'est essentiellement une somme finie).

En particulier, si K est un sous-groupe compact maximal hyperspécial de $G(F)$, si on normalise $\mathrm{d}g$ par $\mathrm{vol}(K, \mathrm{d}g) = 1$ et si $f = 1_{\mathfrak{k}}$ est la fonction caractéristique du $\mathcal{O}_F$-réseau $\mathfrak{k} \subset \mathfrak{g}$ algèbre de Lie de K, on a l'intégrale orbitale $\mathrm{O}_\gamma^G(1_{\mathfrak{k}}, \mathrm{d}g)$. Comme tous les sous-groupes compacts maximaux hyperspéciaux sont conjugués dans $G(F)$, cette intégrale orbitale ne dépend pas du choix de K et on la note simplement O_γ^G dans la suite.

Exemple 3.1 (Suite de 2.1). Soient $(E_i)_{i\in I}$ une famille finie d'extensions finies séparables de F que l'on suppose toutes totalement ramifiées pour simplifier l'exposition. On note n_i le degré de E_i sur F, E_i' l'extension composée $E_i' = E_i F'$ de F' et encore τ l'élément non trivial du groupe de Galois $\mathrm{Gal}(E_i'/E_i) \cong \mathrm{Gal}(F'/F)$.

Pour chaque $i \in I$ et chaque $c_i \in E_i^\times$, on considère le F'-espace hermitien (E_i', Φ_{i,c_i}) où

$$\Phi_{i,c_i}(x, y) = \mathrm{Tr}_{E_i'/F'}(c_i x^\tau y).$$

On note $\mu_i(c_i) \in \mathbb{Z}/2\mathbb{Z}$ son discriminant ; si n_i est premier à p, on a simplement

$$\mu_i(c_i) \equiv v_{E_i}(c_i) + n_i - 1 \pmod 2.$$

Pour chaque $c_I = (c_i)_{i\in I} \in E_I^\times$, on considère le F'-espace hermitien

$$(E_I', \Phi_{I,c_I}) = \bigoplus_{i\in I} (E_i', \Phi_{i,c_i}).$$

Son discriminant est la somme

$$\sum_{i\in I} \mu_i(c_i) \in \mathbb{Z}/2\mathbb{Z}$$

des discriminants des Φ_{i,c_i}. On note

$$G_{I,c_I} \subset \mathrm{Res}_{F'/F}\mathrm{Aut}_{F'}(E_I')$$

le F-groupe réductif des automorphismes du F'-espace hermitien (E_I', Φ_{I,c_I}).

Pour chaque $i \in I$ on note T_i le noyau de l'homomorphisme de tores

$$\mathrm{Res}_{E_i'/F}\mathbb{G}_\mathrm{m} \to \mathrm{Res}_{E_i/F}\mathbb{G}_\mathrm{m}$$

qui envoie x sur $x^\tau x$. Le tore $T_I = \prod_{i\in I} T_i$ est anisotrope sur F. Comme le tore $\prod_{i\in I} \mathrm{Res}_{E_i'/F'}\mathbb{G}_\mathrm{m}$ est de manière naturelle un tore maximal dans $\mathrm{Aut}_{F'}(E_I')$, T_I est naturellement plongé dans $\mathrm{Res}_{F'/F}\mathrm{Aut}_{F'}(E_I')$ et on a en fait

$$T_I \subset G_{I,c_I} \subset \mathrm{Res}_{F'/F}\mathrm{Aut}_{F'}(E_I')$$

pour chaque $c_I \in E_I^\times$.

Pour chaque $i \in I$, soit $\gamma_i \in \mathcal{O}_{E_i'} \subset E_i'$ qui engendre E_i' sur F' et qui vérifie

$$\gamma_i^\tau + \gamma_i = 0,$$

de sorte que $\gamma_I = (\gamma_i)_{i\in I}$ est un élément de l'algèbre de Lie $\mathfrak{t}_I = \bigoplus_{i\in I}\{x_i \in E_i' \mid x_i^\tau + x_i = 0\}$ de T_I. Pour chaque $c_I \in E_I^\times$, γ_I est un élément semi-simple de l'algèbre de Lie $\mathfrak{g}_{I,c_I}$ de G_{I,c_I} de centralisateur T_I.

Supposons de plus que pour tous $i \neq j$ dans I les polynômes minimaux $P_i(x)$ et $P_j(x)$ sur F' de γ_i et γ_j sont premiers entres eux. Alors cette classe de conjugaison semi-simple est régulière et l'intégrale orbitale $\mathrm{O}_{\gamma_I}^{G_{I,c_I}}$ admet l'interprétation concrète suivante : c'est le nombre des $\mathcal{O}_{F'}$-réseaux $M \subset E_I'$ vérifiant les deux conditions suivantes :

- $M^{\perp_{c_I}} = M$, où $M^{\perp_{c_I}} = \{x \in E_I' \mid \Phi_{I,c_I}(x, M) \subset \mathcal{O}_{F'}\}$,
- $\gamma_I M \subset M$.

4. κ-Intégrales orbitales

La classe de *conjugaison stable* d'un élément régulier $\gamma \in \mathfrak{k} \subset \mathfrak{g}$, est l'ensemble des $\gamma' \in \mathfrak{g}$ qui sont conjugués à γ dans $\overline{F} \otimes_F \mathfrak{g}$ par un élément de $G(\overline{F})$. Tout $\gamma' = \mathrm{ad}(g^{-1})(\gamma)$ stablement conjugué à γ définit un 1-cocycle

$$\sigma \mapsto g^{-1}\sigma(g),\ \mathrm{Gal}(\overline{F}/F) \to T(\overline{F}),$$

dont la classe dans $H^1(F,T) := H^1(\mathrm{Gal}(\overline{F}/F), T(\overline{F}))$ ne dépend que de la classe de $G(F)$-conjugaison de γ'. Par construction l'image de cette classe dans l'ensemble pointé $H^1(F,G) := H^1(\mathrm{Gal}(\overline{F}/F), G(\overline{F}))$ est triviale. On construit ainsi une bijection de l'ensemble des classes de $G(F)$-conjugaison dans la classe de conjugaison stable de γ (ensemble pointé par la classe de conjugaison de γ) sur le groupe fini $\mathrm{Ker}(H^1(F,T) \to H^1(F,G))$.

Le centralisateur d'un élément $\gamma' = \mathrm{ad}(g)(\gamma) \in \mathfrak{g}$ stablement conjugué à γ est le tore maximal $T' = gTg^{-1} \subset G$ stablement conjugué à T ; il est donc canoniquement isomorphe à T par l'isomorphisme $T \to T' = gTg^{-1}$ qui envoie t sur gtg^{-1},

isomorphisme qui est en fait défini sur F. Pour tout caractère $\kappa : \mathrm{Ker}(H^1(F,T) \to H^1(F,G)) \to \mathbb{C}^\times$ vu comme une fonction sur la classe de conjugaison stable de γ, on peut former la *κ-intégrale orbitale* en γ d'une fonction $f : \mathfrak{g} \to \mathbb{C}$ localement constante à support compact

$$\mathrm{O}_\gamma^{G,\kappa}(f, \mathrm{d}g) = \sum_{\gamma'} \kappa(\gamma') \mathrm{O}_{\gamma'}^G(f, \mathrm{d}g)$$

où γ' parcourt un système de représentants des classes de conjugaison dans la classe de conjugaison stable de γ. Pour le caractère trivial $\kappa = 1$ on note

$$\mathrm{SO}_{[\gamma]}^G(f, \mathrm{d}g) = \mathrm{O}_\gamma^{G,1}(f, \mathrm{d}g)$$

où $[\gamma]$ est la classe de conjugaison stable de γ, et on dit *intégrale orbitale stable* au lieu de 1-*intégrale orbitale*.

Pour $f = 1_{\mathfrak{k}}$ et $\mathrm{vol}(K, \mathrm{d}g) = 1$ où K est un sous-groupe compact maximal hyperspécial de $G(F)$, on note simplement $\mathrm{O}_\gamma^{G,\kappa}$ la κ-intégrale orbitale et $\mathrm{SO}_{[\gamma]}^G$ l'intégrale orbitale stable ; elles ne dépendent pas du choix de K.

Exemple 4.1 (Suite de 3.1). Pour tout $i \in I$ posons

$$c_i^0 = \frac{\varepsilon^{n-1}}{\frac{\mathrm{d}P_i(x)}{\mathrm{d}x}(\gamma_i) P_{I-\{i\}}(\gamma_i)} \in E_i^\times$$

où $\varepsilon \in \mathcal{O}_{F'}^\times$ est n'importe quel élément vérifiant $\varepsilon^\tau = -\varepsilon$ et où $P_J(x) = \prod_{j \in J} P_j(x)$ quel que soit $J \subset I$. On a

$$\mu_i(c_i^0) \equiv \sum_{j \in I-\{i\}} r_{ji} \pmod 2$$

où r_{ji} est la valuation du résultant des deux polynômes $P_i(x)$ et $P_j(x)$ à coefficients dans F'. Le discriminant de Φ_{I,c_I^0} est trivial et le $\mathcal{O}_{F'}$-réseau $\mathcal{O}_{F'}[\gamma] \subset E_I'$ est autodual pour Φ_{I,c_I^0}. On note simplement G_I le groupe unitaire G_{I,c_I^0}, $\mathfrak{g}_I$ son algèbre de Lie et K_I le sous-groupe maximal hyperspécial de G_I qui fixe le $\mathcal{O}_{F'}$-réseau auto-dual $\mathcal{O}_{F'}[\gamma_I]$. On peut si l'on le souhaite identifier (G_I, K_I) à $(\mathrm{U}(n), \mathrm{U}(n,F) \cap \mathrm{GL}(n, \mathcal{O}_{F'}))$ comme dans la section 2.

Pour chaque $i \in I$ fixons une uniformisante ϖ_{E_i} de E_i et donc aussi de E_i'. Notons Λ_I le noyau de l'application somme des coordonnées $\mathbb{Z}^I \to \mathbb{Z}$ et pour chaque $\lambda \in \Lambda_I$ posons $\varpi_{E_I}^{-\lambda} = (\varpi_{E_i}^{-\lambda_i})_{i \in I}$ et $c_I^\lambda = \varpi_{E_I}^{-\lambda} c_I^0$.

On a l'élément γ_I dans $\mathfrak{t}_I \subset \mathfrak{g}_I$. Pour chaque $\lambda \in \Lambda_I$, le discriminant de la forme hermitienne Φ_{I,c_I^λ} est trivial et il existe donc $x^\lambda \in \mathrm{Aut}_{F'}(E_I')(F')$ tel que $x^\lambda G_{I,c_I^\lambda}(x^\lambda)^{-1} = G_I$. On choisit un tel x^λ et on note γ_I^λ l'élément $\mathrm{ad}(x^\lambda)(\gamma_I)$ de $\mathfrak{g}_I$; la classe de conjugaison dans $\mathfrak{g}_I$ de γ_I^λ ne dépend pas du choix de x^λ (pour $\lambda = 0$ on peut bien sûr prendre $x^\lambda = 1$).

Pour tous $\lambda, \lambda' \in \Lambda_I$, les éléments γ_I^{λ} et $\gamma_I^{\lambda'}$ sont automatiquement stablement conjugués dans $\mathfrak{g}_I$. Ils sont de plus conjugués dans $\mathfrak{g}_I$ si et seulement si $\lambda \equiv \lambda' \pmod{2\Lambda_I}$.

On a donc un classe de conjugaison stable $[\gamma_I]$ bien définie dans $\mathfrak{g}_I$ et une indexation des classes de conjugaison dans cette classe de conjugaison stable par $\Lambda_I/2\Lambda_I$.

On considère l'ensemble $\mathcal{X}_{\gamma_I}$ des $\mathcal{O}_{F'}$-réseaux $M \subset E_I'$ tels qu'il existe $\lambda \in \Lambda_I$ avec

$$M^{\perp_{c_I^{\lambda}}} = M$$

ou ce qui revient au même

$$M^{\perp_{c_I^0}} = \varpi_{E_I'}^{-\lambda} M.$$

Le groupe discret Λ_I agit sur cet ensemble par

$$\lambda \cdot M = \varpi_{E_I'}^{-\lambda} M$$

et on a une application

$$\mu_{\gamma_I} : \mathcal{X}_{\gamma_I}/\Lambda_I \to \Lambda_I/2\Lambda_I$$

qui envoie l'orbite $\Lambda_I \cdot M$ sur la classe de tout λ tel que $M^{\perp_{c_I^0}} = \varpi_{E_I'}^{-\lambda} M$.

Alors, pour tout caractère κ de $\Lambda_I/2\Lambda_I$ la κ-intégrale orbitale $\mathrm{O}_{\gamma_I}^{G_I,\kappa}$ est en fait la somme

$$\sum_{\mu \in \Lambda_I/2\Lambda_I} \kappa(\mu) |\{M \in \mathcal{X}_{\gamma_I} \mid \mu_{\gamma_I}(M) = \mu\}|$$

et en particulier

$$\mathrm{SO}_{\gamma_I}^{G_I} = |\mathcal{X}_{\gamma_I}/\Lambda_I|.$$

5. Dualité de Langlands

Rappelons qu'à tout couple formé d'un groupe réductif connexe G sur un corps séparablement clos k et d'un tore maximal T de G, est associé la *donnée radicielle*

$$(X^*, R, X_*, R^\vee) = (X^*(T), R(G,T), X_*(T), R^\vee(G,T))$$

où X^* (resp. X_*) est le groupe abélien libre de rang fini des caractères (resp. co-caractères) de T et où $R \subset X^*$ (resp. $R^\vee \subset X_*$) est l'ensemble fini des racines (resp. co-racines) de T dans G.

Rappelons aussi que toute donnée radicielle $(X^*, R, X_*, R^\vee)$ formée de deux groupes abéliens libres de rangs finis en dualité et de sous-ensembles finis $R \subset X^*$ et $R^\vee \subset X_*$ vérifiant les axiomes des systèmes de racines et de co-racines, provient d'un couple (G, T) et que le couple (G, T) est uniquement déterminé par sa donnée radicielle à isomorphisme non unique près. Plus précisément, l'homomorphisme du groupe $\mathrm{Aut}(G, T)$ des automorphismes algébriques du couple (G, T) dans celui

$\mathrm{Aut}(X^*, R, X_*, R^\vee)$ des automorphismes de la donnée radicielle induit un isomorphisme

$$\mathrm{Aut}(G,T)/\mathrm{Int}(T) \xrightarrow{\sim} \mathrm{Aut}(X^*, R, X_*, R^\vee)$$

où $\mathrm{Int}\colon G \to \mathrm{Aut}(G)$ est l'homomorphisme de groupes qui associe à un élément de G l'automorphisme intérieur correspondant. L'application $\mathrm{Int}\colon N_G(T) \to \mathrm{Aut}(G,T)$ induit un monomorphisme du groupe de Weyl $W = W(G,T) = N_G(T)/T$ dans $\mathrm{Aut}(G,T)/\mathrm{Int}(T)$ que l'on note encore Int et dont le conoyau est fini.

Partant d'un couple (G,T) sur k comme ci-dessus, il existe donc un couple $(\widehat{G}, \widehat{T})$ formé d'un groupe réductif connexe $\widehat{G}$ sur $\mathbb{C}$ et d'un tore maximal de $\widehat{T}$ de $\widehat{G}$, dont la donnée radicielle est duale de celle de (G,T) au sens où

$$(X^*(\widehat{T}), R(\widehat{G},\widehat{T}), X_*(\widehat{T}), R^\vee(\widehat{G},\widehat{T})) = (X_*(T), R^\vee(G,T), X^*(T), R(G,T)).$$

Ce couple est dit le *dual de Langlands complexe* de (G,T).

Par exemple, on a les groupes duaux suivants (les tores maximaux étant des deux côtés les tores des matrices diagonales)

G	$\widehat{G}$
$\mathrm{GL}(n)$	$\mathrm{GL}(n)$
$\mathrm{PGL}(n)$	$\mathrm{SL}(n)$
$\mathrm{Sp}(2n)$	$\mathrm{SO}(2n+1)$
$\mathrm{SO}(2n)$	$\mathrm{SO}(2n)$

Si G est semi-simple et adjoint, $\widehat{G}$ est semi-simple et simplement connexe, et vice-versa. Plus généralement, si le centre de G est connexe alors le groupe dérivé de $\widehat{G}$ est simplement connexe, et vice-versa.

Si maintenant k est un corps non nécessairement séparablement clos, soit $\bar{k}$ une clôture séparable de k. Pour un groupe réductif connexe G sur k muni d'un tore maximal T sur k, on a comme précédemment la donnée radicielle

$$(X^*, R, X_*, R^\vee) = (X^*(\bar{k}\otimes_k T), R(\bar{k}\otimes_k G, \bar{k}\otimes_k T), X_*(\bar{k}\otimes_k T), R^\vee(\bar{k}\otimes_k G, \bar{k}\otimes_k T)).$$

De plus on a une action de $\mathrm{Gal}(\bar{k}/k)$ sur cette donnée radicielle et donc une action de $\mathrm{Gal}(\bar{k}/k)$ sur la donnée radicielle duale $(X_*, R^\vee, X^*, R)$, d'où un homomorphisme de groupes

$$\rho\colon \mathrm{Gal}(\bar{k}/k) \to \mathrm{Aut}(\widehat{G},\widehat{T})/\mathrm{Int}(\widehat{T}).$$

Pour tout w dans le groupe de Weyl $W = W(\bar{k}\otimes_k G, \bar{k}\otimes_k T) = W(\widehat{G},\widehat{T})$ on peut tordre ρ par w en posant

$${}^w\rho(\sigma) = \mathrm{Int}(w)\rho(\sigma)\mathrm{Int}(w^{-1})$$

quel que soit $\sigma \in \mathrm{Gal}(\bar{k}/k)$. Dire que G est quasi-déployé sur k revient à dire :

(∗) *il existe $w \in W$ tel que ${}^w\rho$ fixe une base Δ du système de racines R ou ce qui revient au même d'une base $\Delta^\vee$ du système de co-racines $R^\vee$.*

La donnée de $(\widehat{G}, \widehat{T})$ munie de l'homomorphisme ρ ci-dessus ne permet pas de retrouver le couple (G, T). Par contre, on a une bijection entre les classes d'isomorphie de couples $(G, [T])$ formés d'un groupe réductif connexe quasi-déployé sur k et d'une classe de conjugaison stable de tores maximaux de G et l'ensemble des classes d'isomorphie de triplets $(\widehat{G}, \widehat{T}, \rho)$ formés d'une groupe réductif connexe sur $\mathbb{C}$, d'un tore maximal $\widehat{T}$ et d'un homomorphisme $\mathrm{Gal}(\bar{k}/k) \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ vérifiant $(*)$.

Exemple 5.1 (Suite de 4.1). Pour $G = \mathrm{Res}_{E'_I/F'}\mathbb{G}_\mathrm{m}$ et $T = \mathrm{Res}_{E'_I/F'}\mathbb{G}_\mathrm{m}$, on a $\widehat{G} = \mathrm{Aut}(\mathbb{C}^{\mathrm{Hom}_{F'}(E'_I, \bar{F})})$, $\widehat{T} = (\mathbb{C}^\times)^{\mathrm{Hom}_{F'}(E'_I, \bar{F})}$ et l'homomorphisme $\rho\colon \mathrm{Gal}(\bar{F}/F') \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ envoie σ sur l'élément du groupe de Weyl

$$W(\widehat{G}, \widehat{T}) = \mathrm{Int}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T}) \subset \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$$

défini par la permutation de $\mathrm{Hom}_{F'}(E'_I, \bar{F})$ induite par l'action de σ sur $\bar{F}$.

Dans ce cas $\mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ est le produit semi-direct de $W(\widehat{G}, \widehat{T})$ par l'automorphisme extérieur de $\widehat{G}$ donné par $g \to {}^{\mathrm{t}}g^{-1}$ où la transposition est relative à la base canonique de $\mathbb{C}^{\mathrm{Hom}_{F'}(E'_I, \bar{F})}$.

Pour $G = G_I$ et $T = T_I$, on a $\widehat{G} = \mathrm{Aut}(\mathbb{C}^{\mathrm{Hom}_F(E_I, \bar{F})})$, $\widehat{T} = (\mathbb{C}^\times)^{\mathrm{Hom}_F(E_I, \bar{F})}$ et l'homomorphisme $\rho\colon \mathrm{Gal}(\bar{F}/F) \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ envoie $\sigma \in \mathrm{Gal}(\bar{F}/F') \subset \mathrm{Gal}(\bar{F}/F)$ sur l'élément du groupe de Weyl $W(\widehat{G}, \widehat{T})$ défini par la permutation de $\mathrm{Hom}_F(E_I, \bar{F}) = \mathrm{Hom}_{F'}(E'_I, \bar{F})$ induite par l'action de σ sur $\bar{F}$, et il envoie $\sigma \in \mathrm{Gal}(\bar{F}/F) - \mathrm{Gal}(\bar{F}/F')$ sur le produit de ce même élément du groupe de Weyl et de l'automorphisme extérieur ci-dessus.

Pour le groupe unitaire quasi-déployé $G = \mathrm{U}(n)$ et T le tore maximal le plus déployé formé des matrices unitaires diagonales, on a $\widehat{G} = \mathrm{GL}(n, \mathbb{C})$, $\widehat{T}$ est le tore des matrices diagonales de $\mathrm{GL}(n, \mathbb{C})$ et l'homomorphisme $\rho\colon \mathrm{Gal}(\bar{F}/F) \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ se factorise par $\mathrm{Gal}(F'/F)$ et envoie l'élément non trivial τ sur l'automorphisme extérieur $g \mapsto \Phi_n {}^{\mathrm{t}}g^{-1} \Phi_n$.

6. Groupes endoscopiques

Soit G un groupe réductif connexe quasi-déployé sur le corps local non archimédien F muni d'un tore maximal T.

Soit $(\widehat{G}, \widehat{T})$ le groupe dual de $(\bar{F} \otimes_F G, \bar{F} \otimes_F T)$ muni de l'homomorphisme $\rho\colon \mathrm{Gal}(\bar{F}/F) \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$.

La dualité de Tate-Nakayama identifie ce groupe fini $\pi_0(\widehat{T}^{\mathrm{Gal}(\bar{F}/F)})$ au dual de Pontryagin de $H^1(F, T)$. Plus généralement la dualité de Tate-Nakayama étendue par Kottwitz identifie $H^1(F, G)$ au groupe fini $\pi_0(Z(\widehat{G})^{\mathrm{Gal}(\bar{F}/F)})$ où on a noté $Z(\widehat{G})$ le centre de $\widehat{G}$. Par suite, le dual du noyau $\mathrm{Ker}(H^1(F, T) \to H^1(F, G))$ est le quotient $\pi_0(\widehat{T}^{\mathrm{Gal}(\bar{F}/F)})/\pi_0(Z(\widehat{G})^{\mathrm{Gal}(\bar{F}/F)})$.

Pour simplifier l'exposition on suppose dans la suite que T est anisotrope de sorte que $\widehat{T}^{\mathrm{Gal}(\overline{F}/F)}$ est fini, tout comme $Z(\widehat{G})^{\mathrm{Gal}(\overline{F}/F)} \subset \widehat{T}^{\mathrm{Gal}(\overline{F}/F)}$.

Si $s \in \widehat{T}^{\mathrm{Gal}(\overline{F}/F)}$, la composante neutre $\widehat{H}$ de son centralisateur dans $\widehat{G}$ est un groupe réductif connexe qui contient le tore maximal $\widehat{T}$. La donnée radicielle de $(\widehat{H}, \widehat{T})$, qui est contenue dans celle de $(\widehat{G}, \widehat{T})$, est stable sous l'action de $\mathrm{Gal}(\overline{F}/F)$. L'homomorphisme $\rho\colon \mathrm{Gal}(\overline{F}/F) \to \mathrm{Aut}(\widehat{G}, \widehat{T})/\mathrm{Int}(\widehat{T})$ a son image contenue dans $\mathrm{Aut}(\widehat{G}, \widehat{H}, \widehat{T})/\mathrm{Int}(\widehat{T})$ et induit donc un homomorphisme

$$\rho_H\colon \mathrm{Gal}(\overline{F}/F) \to \mathrm{Aut}(\widehat{H}, \widehat{T})/\mathrm{Int}(\widehat{T}).$$

Par suite il existe un couple (unique à isomorphisme près) formé d'un groupe réductif connexe quasi-déployé H sur F et d'une classe de conjugaison stable $[T]$ de tores maximaux de H, de dual $(\widehat{H}, \widehat{T})$ avec l'homomorphisme ρ_H. Un tel couple est dit *endoscopique* pour (G, T). Tout tore S dans $[T]$ est muni d'un isomorphisme naturel avec T (ce qui justifie notre notation) et est donc anisotrope.

Exemple 6.1 (Suite de 5.1). Pour $G = G_I$ et $T = T_I$ on a T_I anisotrope et $\widehat{T}^{\mathrm{Gal}(\overline{F}/F)} = \{\pm 1\}^I \subset (\mathbb{C}^\times)^I \subset \prod_{i\in I}(\mathbb{C}^\times)^{\mathrm{Hom}_F(E_i,\overline{F})} = \widehat{T}$. Tout $s = (s_i)_{i\in I} \in \widehat{T}^{\mathrm{Gal}(\overline{F}/F)}$ définit une partition $I = I_1 \amalg I_2$ où I_1 (resp. I_2) est l'ensemble des $i \in I$ tel que $s_i = 1$ (resp. $s_i = -1$). Le groupe endoscopique défini par s est le produit de groupes unitaires $G_{I_1} \times_F G_{I_2}$, muni de la classe de conjugaison stable de tores maximaux $[T_{I_1} \times_F T_{I_2}]$.

7. Lemme Fondamental

Conjecture 7.1 (Langlands-Shelstad [20]). Soient G un groupe réductif non ramifié sur F, T un tore maximal anisotrope de G, $\gamma \in \mathfrak{t}$ un élément régulier semi-simple dans $\mathfrak{g}$ comme dans la section 3, et $s \in \widehat{T}^{\mathrm{Gal}(\overline{F}/F)}$ définissant un caractère κ de $\mathrm{Ker}(H^1(F, T) \to H^1(F, G))$ par la dualité de Tate-Nakayama comme dans la section précédente. On a la κ-intégrale orbitale $\mathrm{O}_\gamma^{G,\kappa}$, le groupe endoscopique H muni de sa classe de conjugaison stable de tore maximaux $[T]$ défini par s et l'*intégrale orbitale stable* $\mathrm{SO}^H_{[\gamma]}$.

Alors

$$\mathrm{O}_\gamma^{G,\kappa} = \varepsilon(\gamma)\mathrm{D}_H^G(\gamma)\mathrm{SO}^H_{[\gamma]}.$$

où $\varepsilon(\gamma)$ est une racine de l'unité qui est définie précisément dans [20] et où le discriminant $\mathrm{D}_H^G(\gamma)$ est défini par

$$\mathrm{D}_H^G(\gamma) = \prod_{\alpha\in R(G,T)-R(H,T)} |\alpha(\gamma)|_F^{1/2}.$$

On a noté $|a|_{\overline{F}} = q^{-v_F(a)}$ pour tout $a \in \overline{F}^\times$ où q est de nombre des élément du corps résiduel de F et $v_F\colon \overline{F}^\times \to \mathbb{Q}$ est la valuation normalisée par $v_F(\varpi_F) = 1$ pour toute uniformisante ϖ_F de F.

Remarque 7.2. (i) La racine de l'unité $\varepsilon(\gamma)$ a été calculée par Waldspurger pour les groupes classiques dans [29].

(ii) Kottwitz a montré comment pointer la classe de conjugaison stable de γ dans G à l'aide de la section de Kostant pour simplifier la définition de Langlands-Shelstad de $\varepsilon(\gamma)$ (cf. [15]).

(iii) D'après Waldspurger (cf. [30]), il suffit de démontrer le Lemme Fondamental pour F d'égales caractéristiques, c'est-à-dire $F = \mathbb{F}_q((t))$ pour un corps fini $\mathbb{F}_q$.

8. Résultats

Outre le cas particulier de SL(2) qui est à l'origine du Lemme Fondamental (cf. [18]) plusieurs cas particuliers de 7.1 ont déjà été démontrés sans recours à la géométrie algébrique : le cas de SL(n) par Waldspurger (cf. [27]), le cas de U(3) par Kottwitz d'une part (cf. [14]) et par Rogawski d'autre part (cf. [25]), et le cas de Sp(4) par Hales d'une part (cf. [9]) et Weissauer d'autre part (cf. [31]).

Par contre les deux théorèmes ci-dessous sont obtenus par voie géométrique, ce qui suppose que F soit d'égales caractéristiques (voir cependant la remarque 7.2 (iii)).

Théorème 8.1 (Goresky, Kottwitz et MacPherson [7]). *En plus des hypothèse de la conjecture* 7.1*, supposons que le tore anisotrope T est non ramifié, c'est-à-dire qu'il se déploie sur une extension finie non ramifiée de F, et que l'élément régulier semi-simple $\gamma \in \mathfrak{t}$ est d'égales valuations, c'est-à-dire que la valuation de $\alpha'(\gamma)$ ne dépende pas de la racine α de T dans G (α' est la dérivée de α). Alors*

$$\mathrm{O}_\gamma^{G,\kappa} = \varepsilon(\gamma)\mathrm{D}_H^G(\gamma)\mathrm{SO}_{[\gamma]}^H.$$

Théorème 8.2 (Laumon et Ngô [22]). *Reprenons les notations et les hypothèses de l'exemple* 4.1*. Nous avons donc le groupe réductif quasi-déployé $G = G_I$, une classe de conjugaison stable $[\gamma]$ dans $\mathfrak{g}_I$ pointée par un élément $\gamma = \gamma_I$ dans l'algèbre de Lie et un tore maximal anisotrope $T = T_I$ de G.*

Fixons une partition $I = I_1 \sqcup I_2$, notons $H = G_{I_1} \times G_{I_2}$ le groupe endoscopique correspondant et encore $[\gamma]$ la classe de conjugaison stable dans $\mathfrak{h}$ de $\gamma \in \mathfrak{t}_I = \mathfrak{t}_{I_1} \oplus \mathfrak{t}_{I_2} \subset \mathfrak{h}$.

Supposons que F soit d'égales caractéristiques $p > n$. Alors on a la relation

$$\mathrm{O}_\gamma^{\kappa} = (-1)^r q^r \, \mathrm{SO}_{[\gamma]}^H$$

où

$$r = r_{I_1,I_2} = \sum_{\substack{i_1 \in I_1 \\ i_2 \in I_2}} r_{i_1,i_2}.$$

(*On rappelle que $r_{ij} = r_{ji}$ est la valuation du résultant* $\mathrm{Res}(P_i, P_j) \in F'$ *des polynômes minimaux $P_i(x)$ et $P_j(x)$ de γ_i et γ_j.*)

9. Fibres de Springer affines

Les preuves des théorèmes 8.1 et 8.2 sont fondées sur une interprétation cohomologique des deux membres du Lemme Fondamental à l'aide de la formule des points fixes de Grothendieck-Lefschetz. Cette interprétation fait intervenir les fibres de Springer affines qui sont des analogues pour les groupes de lacets des fibres de Springer classiques.

Soient k un corps algébriquement clos et G un groupe réductif sur k. On note $G((t))$ le ind-k-schéma en groupes de lacets des $k((t))$-points de G et $G[[t]] \subset G((t))$ son sous-k-schéma en groupes des $k[[t]]$-points de G. On ne tient pas compte ici des nilpotents éventuels de ces schémas.

Le ind-k-schéma réduit quotient $X = G((t))/G[[t]]$ est appelé la *grassmannienne affine* de G ; il est réunion croissante de k-schémas projectifs. Le ind-k-schéma en groupes $G((t))$ agit par translation à gauche sur X.

Pour $G = \mathrm{GL}_k(n)$ cette grassmannienne affine n'est autre que le ind-k-schéma réduit des $k[[t]]$-réseaux dans $k((t))^n$, c'est-à-dire des sous-$k[[t]]$-modules $M \subset k((t))^n$ tels qu'il existe un entier $N \geq 0$ avec $t^N k[[t]]^n \subset M \subset t^{-N} k[[t]]^n \subset k((t))^n$.

Soit $\mathfrak{g}$ l'algèbre de Lie de G et γ un élément régulier semi-simple de $\mathfrak{g}((t)) = \mathfrak{g} \otimes_k k((t))$. La fibre de Springer affine en γ est le sous-ind-k-schéma fermé réduit X_γ de X dont l'ensemble des k-points est

$$X_\gamma(k) = \{gG(k[[t]]) \in X(k) \mid \mathrm{ad}(g^{-1})(\gamma) \in \mathfrak{g}[[t]]\}.$$

Pour que cet ensemble soit non vide il faut que, à conjugaison près on ait $\gamma \in \mathfrak{g}[[t]] = \mathfrak{g} \otimes_k k[[t]]$, ce que nous supposons dans la suite.

Pour $\mathrm{GL}_k(n)$ et $\gamma \in \mathrm{gl}(n, k[[t]])$ une matrice régulière semi-simple, la fibre de Springer affine est encore le ind-k-schéma réduit des $k[[t]]$-réseaux $M \subset k((t))^n$ tels que $\gamma M \subset M$.

Le centralisateur T de γ dans $G((t))$ est un tore maximal par hypothèse. L'action de $T \subset G((t))$ par translation à gauche sur X respecte le fermé X_γ. Le tore T contient un plus grand sous-tore de la forme $A((t))$ pour A un sous-tore de $\mathfrak{g}$ sur k (le sous-tore déployé maximal). Le groupe des co-caractères $X_*(A)$ agit librement sur X_γ par

$$\lambda \cdot gG[[t]] = \lambda(t) gG[[t]].$$

Dans [11], Kazhdan et Lusztig ont montré que X_γ est en fait un schéma localement de type fini et de dimension finie sur k, et que le quotient de X_γ par l'action libre de $X_*(A)$ ci-dessus est un k-schéma projectif, l'application quotient $X_\gamma \to X_\gamma/X_*(A)$ étant un revêtement étale galoisien de groupe de Galois $X_*(A)$.

La structure de ind-k-schéma de la grassmannienne affine X induit sur la fibre de Springer affine X_γ une structure analogue : X_γ est une réunion croissante de sous-k-schémas fermés projectifs.

Soit ℓ un nombre premier distinct de la caractéristique de k. On peut définir la cohomologie étale ℓ-adique de X_γ

$$H^i(X_\gamma, \mathbb{Q}_\ell).$$

comme la limite projective des cohomologies étales ℓ-adiques des sous-k-schémas fermés projectifs ci-dessus.

Si $k = \overline{\mathbb{F}}_q$ est la clôture algébrique d'un corps fini $\mathbb{F}_q$ et si G et γ sont définis sur ce corps fini, il en est bien entendu de même de $G((t))$, $G[[t]]$, X et X_γ. Par suite, pour tout nombre premier ℓ distinct de la caractéristique de k on a une action de l'endomorphisme de Frobenius géométrique Frob_q relatif à $\mathbb{F}_q$ sur la cohomologie étale ℓ-adique $H^i(X_\gamma, \mathbb{Q}_\ell)$.

Rappelons que si Z est un k-schéma propre sur $k = \overline{\mathbb{F}}_q$, qui est défini que $\mathbb{F}_q$, Deligne a montré que, quel que soit l'entier i, chaque valeur propre λ de Frob_q sur $H^i(Z, \mathbb{Q}_\ell)$ est un entier algébrique de poids $\leq i$, c'est-à-dire qu'il existe un entier $w(\lambda) \leq i$ tel que $|\iota(\lambda)| = q^{-w(\lambda)/2}$ pour tout plongement de $\iota\colon \mathbb{Q}(\lambda) \hookrightarrow \mathbb{C}$. On dit que la cohomologie étale ℓ-adique de Z est *pure* si pour tout entier i et toute valeur propre λ de Frob_q sur $H^i(Z, \mathbb{Q}_\ell)$, $w(\lambda) = i$ (cf. [6]). Toujours d'après Deligne c'est le cas si Z est supposé de plus lisse sur k (cf. loc. cit.). Les fibres de Springer sont des exemples de schémas propres non lisses dont la cohomologie étale ℓ-adique est pure (cf. [26]).

Cette notion de pureté garde un sens pour les fibres de Springer affines X_γ et Goresky, Kottwitz et MacPherson ont conjecturé dans [7] :

Conjecture 9.1. La cohomologie étale ℓ-adique de X_γ est pure.

Ils ont démontré cette conjecture dans un cas particulier (cf. [8]) :

Théorème 9.2 (Goresky, Kottwitz et MacPherson). *Supposons que γ est d'égales valuations, c'est-à-dire que la valuation de $\alpha'(\gamma)$ ne dépend pas de la racine α de T dans G ($\alpha'\colon \mathfrak{t} \to k((t))$ est la dérivée de α). Alors X_γ peut être pavé par des fibrés en espaces affines standard sur des variétés projectives et lisses sur k. En particulier la cohomologie étale ℓ-adique de X_γ est pure.*

10. L'approche de Goresky, Kottwitz et MacPherson

Soient toujours k une clôture algébrique d'un corps fini $\mathbb{F}_q$ et G un groupe réductif sur k que l'on suppose défini sur $\mathbb{F}_q$.

Soit $T \subset G$ un tore maximal défini sur $\mathbb{F}_q$ et $\gamma \in \mathfrak{t}[[t]]$ un élément régulier (semi-simple) dans $\mathfrak{g}((t))$ que l'on suppose rationnel sur $\mathbb{F}_q$, où bien sûr $\mathfrak{t}$ et $\mathfrak{g}$ sont les algèbres de Lie de T et G. Comme dans la section précédente on a la fibre de Springer affine

$$X_\gamma = \{gG[[t]] \mid g^{-1}\gamma g \in \mathfrak{g}[[t]]\} \subset X$$

en γ qui est laissée globalement invariante par l'action par translations à gauche du centralisateur $T((t)) \subset G((t))$ de γ. En particulier on a une action du tore $T \subset T((t))$ et du groupe des co-caractères $\Lambda = X_*(T) \cong t^{X_*(T)} \subset G((t))$ sur X qui laissent globalement invariante X_γ ; l'action de Λ sur X_γ est libre et commute à celle de T. La fibre de Springer X_γ et les actions de T et Λ sont définies sur $\mathbb{F}_q$.

Le lieu des points fixes de T agissant sur X est le fermé $X^T = T((t))/T[[t]] \subset G((t))/G[[t]] = X$ qui est automatiquement contenu dans X_γ et est donc le lieu des points fixes X_γ^T de l'action de T sur X_γ. Ce lieu est bien entendu globalement stable sous l'action de Λ et on peut l'identifier à Λ avec l'action de Λ par translation sur lui-même. La cohomologie étale ℓ-adique de $X_\gamma^T = X^T$ avec l'action induite de Λ est le $\mathbb{Q}_\ell[\Lambda]$-module $\mathbb{Q}_\ell[[\Lambda]]$.

Goresky, Kottwitz et MacPherson donnent une formule explicite pour la cohomologie étale ℓ-adique de X_γ sous l'hypothèse que cette cohomologie est pure. Ils déduisent cette formule d'une formule explicite pour la cohomologie étale ℓ-adique T-équivariante de X_γ qu'ils établissent en premier.

Avant d'énoncer leur résultat rappelons que la cohomologie étale ℓ-adique T-équivariante de $\mathrm{Spec}(k)$ avec l'action triviale de T est la $\mathbb{Q}_\ell$-algèbre graduée

$$H_T^*(\mathrm{Spec}(k), \mathbb{Q}_\ell) = \bigoplus_{n\geq 0} H_T^{2n}(\mathrm{Spec}(k), \mathbb{Q}_\ell) = \bigoplus_{n\geq 0} \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell,$$

où $\chi \in X^*(T)$ correspond à la classe de Chern du fibré en droites $\mathcal{L}_\chi$ sur le champ algébrique $[\mathrm{Spec}(k)/T]$, obtenu en poussant par χ le T-torseur universel. Rappelons aussi que cette $\mathbb{Q}_\ell$-algèbre graduée agit par le cup-produit sur la cohomologie étale ℓ-adique T-équivariante de tout ind-k-schéma muni d'une action de T. On a donc

$$H_T^*(X_\gamma^T, \mathbb{Q}_\ell) = \bigoplus_{n\geq 0} \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[[\Lambda]]$$

en tant que $\bigoplus_{n\geq 0} \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[\Lambda]$-module gradué et un homomorphisme de $\bigoplus_{n\geq 0} \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[\Lambda]$-modules graduées

$$H_T^*(X_\gamma, \mathbb{Q}_\ell) \to H_T^*(X_\gamma^T, \mathbb{Q}_\ell) = \bigoplus_{n\geq 0} \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[[\Lambda]].$$

de restriction au lieu des points fixes sous T. On notera par la multiplication à gauche l'action de $\mathrm{Sym}^*(X^*(T)(-1))$ et par la multiplication à droite celle de $\mathbb{Q}_\ell[\Lambda]$ pour clarifier l'exposition.

Théorème 10.1 (Goresky, Kottwitz et MacPherson). *Supposons que la cohomologie étale ℓ-adique de X_γ soit pure. Alors la flèche de restriction ci-dessus est injective et son image est formée en degré n des $f \in \mathrm{Sym}^n(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[[\Lambda]]$ tels que, pour toute racine $\alpha \in X^*(T)$ de T dans G et tout entier $d = 1, 2, \ldots, v(\alpha(\gamma))$, on ait*

$$f(1-\alpha^\vee)^d \in \alpha^d \mathrm{Sym}^{n-d}(X^*(T)(-1)) \otimes \mathbb{Q}_\ell[[\Lambda]]$$

où $\alpha' : \mathfrak{t}((t)) \to k((t))$ est la dérivée de α et où $\alpha^\vee \in \Lambda$ est la co-racine correspondante.

Le théorème 8.1 résulte des théorèmes 9.2 et 10.1

11. Notre approche avec Ngô

L'idée de départ est d'essayer de déformer les fibres de Springer en espérant qu'après déformation les choses deviennent plus simples. Bien sûr nous sommes guidés par l'analogie avec la résolution simultanée de Grothendieck-Springer qui regroupe en une seule famille toutes les fibres de Springer classiques.

Les fibres de Springer affines se comportent mal du point de vue des déformations et le point clé de notre approche est de les remplacer par des objets de nature globale que sont les jacobiennes compactifiées de courbes singulières, et de manière plus proche de la théorie des groupes par les fibres du morphisme de Hitchin.

Pour rendre les choses plus transparente je me limite dans la suite au cas où $G = \mathrm{GL}_k(n)$ bien que cela ne soit pas suffisant pour notre preuve du théorème 8.2.

Se donner un élément régulier semi-simple γ de $\mathfrak{g}((t))$ équivaut à se donner un polynôme unitaire séparable $P(x) \in k((t))[x]$ de degré n, et une base du $k((t))$-espace vectoriel $k((t))[x]/(P(x))$ de dimension n. Pour que la fibre de Springer affine X_γ soit non vide il faut que $P(x) \in k[[t]][x]$; pour simplifier l'exposition on suppose dans la suite que c'est bien le cas et que $P(0) \in tk[[t]]$. On introduit alors l'anneau $R_\gamma = k[[t]][\gamma] = k[[t]][x]/(P(x))$; c'est une k-algèbre réduite, non normale, d'anneau total des fractions $\mathrm{Frac}(R_\gamma) = k((t))[x]/(P(x))$.

Pour toute k-algèbre réduite complète R isomorphe à $k[[x, y]]/(f)$ pour $f \in (x, y) \subset k[[x, y]]$ (comme l'est R_γ), on dispose d'un k-schéma en groupes commutatifs P_R et d'une "compactification" équivariante $\bar{P}_R$ de ce dernier.

Le groupe de k-points de P_R est simplement $\mathrm{Frac}(R)^\times / R^\times$ où $\mathrm{Frac}(R)$ est l'anneau total des fractions de R ; le groupe des composantes connexes de P_R est $\mathbb{Z}^I$ où I est l'ensemble des points génériques de $\mathrm{Spec}(R)$ (ou ce qui revient au même des facteurs irréductibles de f) et la composante neutre P_R^0 de P_R est un k-schéma en groupes de type fini extension d'une tore $\mathbb{G}_{\mathrm{m},k}^I / \mathbb{G}_{\mathrm{m},k}$ par un groupe unipotent. On peut voir P_R comme l'espace de modules des R-modules M libres de rang 1 munis d'un isomorphisme $\mathrm{Frac}(R) \otimes_R M \cong \mathrm{Frac}(R)$

Le k-schéma $\bar{P}_R$ est lui l'espace de modules des R-modules M de type fini sans torsion munis d'un isomorphisme $\mathrm{Frac}(R) \otimes_R M \cong \mathrm{Frac}(R)$; P_R qui est évidemment contenu dans $\bar{P}_R$, agit par produit tensoriel sur ce dernier.

Proposition 11.1 (cf. [21]). *La fibre de Springer affine X_γ est homéomorphe à $\bar{P}_{R_\gamma}$.*

Soit maintenant C une courbe réduite, connexe et projective sur k. Le k-champ algébrique de Picard $\mathrm{Pic}(C)$ de C est l'espace de modules des $\mathcal{O}_C$-Modules inversibles $\mathcal{L}$. Ce champ algébrique est lisse sur k et le produit tensoriel le munit d'une structure de champ algébrique en "groupes commutatifs".

Les composantes connexes de $\mathrm{Pic}(C)$ sont découpées par l'invariant discret $\deg(\mathcal{L}) = \chi(C, \mathcal{L}) - \chi(C, \mathcal{O}_C)$; elles sont de type fini si et seulement si C est irréductible.

Mayer et Mumford ont introduit une "compactification" équivariante de $\mathrm{Pic}(C)$ (cf. [2]). Plus précisément ils ont introduit le k-champ algébrique $\overline{\mathrm{Pic}}(C)$ des $\mathcal{O}_C$-Modules cohérents $\mathcal{F}$ sans torsion de rangs génériques tous égaux à 1. On a une immersion ouverte évidente de $\mathrm{Pic}(C) \hookrightarrow \overline{\mathrm{Pic}}(C)$ et l'action par translation de $\mathrm{Pic}(C)$ sur lui-même se prolonge en l'action de $\mathrm{Pic}(C)$ sur $\overline{\mathrm{Pic}}(C)$ induite par le produit tensoriel. En général $\overline{\mathrm{Pic}}(C)$ n'est pas lisse sur k.

Les composantes connexes de $\mathrm{Pic}(C)$ sont elles aussi découpées par le degré de $\mathcal{F}$. Elles sont de type fini que si et seulement si C est irréductible. Par contre elles ne sont pas irréductibles en général car $\mathrm{Pic}(C)$ n'est pas toujours dense dans $\overline{\mathrm{Pic}}(C)$ et la dimension de $\overline{\mathrm{Pic}}(C)$ peut être strictement plus grande que celle de $\mathrm{Pic}(C)$.

Cependant, si les singularités de C sont toutes planes, c'est-à-dire si le complété formel de $\mathcal{O}_{C,c}$ est isomorphe à $k[[x, y]]/(f(x, y))$ pour $f(x, y) \in (x, y) \subset k[[x, y]]$ quel que soit le point singulier c de C, alors Rego d'une part (cf. [24]) et Altmann, Iarrobino et Kleiman d'autre part (cf. [1]) ont montré que $\mathrm{Pic}(C)$ est dense dans $\overline{\mathrm{Pic}}(C)$ (ces auteurs ne considèrent que le cas où C est irréductible mais leur argument est général).

Supposons dorénavant que notre courbe C réduite, connexe et projective sur k a toutes ses singularités planes. Notons $C^{\mathrm{sing}} \subset C$ l'ensemble fini de ses singularités. On a un k-morphisme de champs algébriques

$$\prod_{c \in C^{\mathrm{sing}}} \bar{P}_{\widehat{\mathcal{O}}_{C,c}} \to \overline{\mathrm{Pic}}(C)$$

qui envoie $(M_c)_c$ sur le $\mathcal{O}_C$-Module cohérent obtenu en recollant $\mathcal{O}_{C-C^{\mathrm{sing}}}$ et les M_c le long de $(C - C^{\mathrm{sing}}) \cap \coprod_{c \in C^{\mathrm{sing}}} \mathrm{Spec}(\widehat{\mathcal{O}}_{C,c}) = \coprod_{c \in C^{\mathrm{sing}}} \mathrm{Spec}(\mathrm{Frac}(\widehat{\mathcal{O}}_{C,c}))$. Ce morphisme induit un morphisme de champs algébriques en groupes commutatifs

$$\prod_{c \in C^{\mathrm{sing}}} P_{\widehat{\mathcal{O}}_{C,c}} \to \mathrm{Pic}(C).$$

Proposition 11.2. *Le k-morphisme de champs algébriques déduit des morphismes précédents par passage au quotient*

$$\prod_{c \in C^{\mathrm{sing}}} [\bar{P}_{\widehat{\mathcal{O}}_{C,c}} / P_{\widehat{\mathcal{O}}_{C,c}}] \to [\overline{\mathrm{Pic}}(C)/\mathrm{Pic}(C)]$$

est un isomorphisme.

Compte-tenu de cette proposition et de la proposition 11.1 la conjecture de pureté de Goresky, Kottwitz et MacPherson 9.1 implique en particulier la conjecture suivante.

Conjecture 11.3. Soit C une courbe intègre et projective sur la clôture algébrique d'un corps fini. On suppose que toutes les singularités de C sont planes et unibranches. Alors la cohomologie ℓ-adique de $\overline{\mathrm{Pic}}(C)^0$ est pure.

Les propositions 11.1 et 11.2 permettent de donner une interprétation cohomologique des intégrales orbitales pour G en termes de $\overline{\mathrm{Pic}}(C)$ pour une courbe C convenable (cf. [21]). Un des gros avantages des champs de Picard compactifiés par rapport aux fibres de Springer affines est que toute déformation de C donne lieu à une déformation de $\overline{\mathrm{Pic}}(C)$.

La fibration de Hitchin fournit de manière naturelle du point de vue de la théorie des groupes une telle déformation (cf. [23]).

Changeons de notations. Soit maintenant C une courbe connexe, projective et lisse de genre $g \geq 2$ sur un corps k algébriquement clos et D un diviseur effectif de degré $\geq 2g-2$ sur C. On dispose du k-champ algébrique $\mathcal{M}$ des couples $(\mathcal{E}, \theta)$ où $\mathcal{E}$ est un fibré vectoriel de rang n sur C et $\theta\colon \mathcal{E} \to \mathcal{E}(D)$ est un endomorphisme tordu de $\mathcal{E}$. Soit $\mathcal{A}$ le k-espace vectoriel de dimension finie $\bigoplus_{i=1}^{n} H^0(X, \mathcal{O}_C(iD))$ vu comme un k-schéma affine. La fibration de Hitchin (cf. [10])

$$m\colon \mathcal{M} \to \mathcal{A}$$

est le morphisme qui envoie $(\mathcal{E}, \theta)$ sur le polynôme caractéristique de θ, c'est-à-dire défini par

$$m(\mathcal{E}, \theta) = (-\mathrm{tr}(\theta), \mathrm{tr}(\wedge^2\theta), \ldots, (-1)^n \mathrm{tr}(\wedge^n\theta))$$

Tout point $a \in \mathcal{A}$ définit un diviseur de Cartier C_a dans la surface réglée $\mathbb{V}(\mathcal{O}_C(-D))$, diviseur appelé par Hitchin la *courbe spectrale* en a, à savoir le diviseur d'équation

$$u^n + p^*a_1u^{n-1} + \cdots + p^*a_n = 0$$

où $p\colon \mathbb{V}(\mathcal{O}_C(-D)) \to C$ est la projection canonique et u est la section universelle de $p^*\mathcal{O}(-D)$. La restriction π_a de p à cette courbe C_a en fait un revêtement fini ramifié de C. Par construction $\pi_{a,*}\mathcal{O}_{C_a}$ est isomorphe à

$$\mathrm{Sym}_{\mathcal{O}_C}(\mathcal{O}_C(-D))/\mathcal{I}_a$$

où $\mathcal{I}_a$ est l'Idéal engendré par l'image de l'homomorphisme

$$\mathcal{O}_C(-nD) \to \bigoplus_{i=0}^{n} \mathcal{O}_C(-iD) \subset \mathrm{Sym}_{\mathcal{O}_C}(\mathcal{O}_C(-D))$$

de composantes $(a_n, a_{n-1}, \ldots, a_1, 1)$.

Supposons que C_a est réduite. On a alors un morphisme du champ algébrique $\overline{\mathrm{Pic}}(C_a)$ des $\mathcal{O}_{C_a}$-Modules cohérents sans torsion de rang 1 en tout point générique de C_a dans la fibre de la fibration de Hitchin $m^{-1}(a) = \mathcal{M}_a$ en a :

$$\overline{\mathrm{Pic}}(C_a) \to m^{-1}(a),\ \mathcal{F} \mapsto (\mathcal{E}, \theta)$$

où $\mathcal{E} = \pi_{a,*}\mathcal{F}$ et

$$\theta\colon \mathcal{O}_C(-D) \subset \mathrm{Sym}_{\mathcal{O}_C}(\mathcal{O}_C(-D))/\mathcal{I}_a = \pi_{a,*}\mathcal{O}_{C_a} \to \mathcal{E}nd_{\mathcal{O}_C}(\mathcal{E}).$$

Proposition 11.4 (Beauville, Narasimhan et Ramanan [4]). *Le morphisme de champs algébriques ci-dessus est un isomorphisme.*

Compte tenu des proposition 11.1 et 11.2, on a donc un lien entre les fibres de la fibration de Hitchin et fibres de Springer affines via les jacobiennes compactifiées. La fibration de Hitchin peut donc être vue comme une famille naturelle de (produits de) fibres de Springer affines. Elle joue un peu le rôle de la résolution simultanée de Grothendieck-Springer pour les fibres de Springer affines. C'est le point de départ de notre démonstration du théorème 8.2.

Remerciements. Je remercie Pierre-Henri Chaudouard et Ngô Bao Châu pour leur aide durant la préparation de cet exposé.

Références

[1] Altman, A., Iarrobino, A., et Kleiman, S., Irreducibility of the Compactified Jacobian. Dans *Real and complex singularities, Oslo 1976* (ed. par. P. Holm), Sijthoff and Noordhoff, Alphen aan den Rijn 1977, 1–12.

[2] Altman, A., et Kleiman, S., Compactifying the Jacobian. *Bull. Amer. Math. Soc.* **82** (1976), 947–949.

[3] Arthur, J., Toward a stable trace formula. Dans *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 507–517.

[4] Beauville, A., Narasimhan, M., et Ramanan, S., Spectral curve and the generalised theta divisor. *J. Reine Angew. Math.* **398** (1989), 169–179.

[5] Dat, J.-F., Lemme fondamental et endoscopie, une approche géométrique. *Séminaire Bourbaki* **940** (2004).

[6] Deligne, P., La conjecture de Weil. II. *Inst. Hautes Études Sci. Publ. Math.* **52** (1980), 313–428.

[7] Goresky, M. , Kottwitz, R., et MacPherson, R., Homology of affine Springer fiber in the unramified case. *Duke Math. J.* **121** (2004), 509–561.

[8] Goresky, M. , Kottwitz, R., et MacPherson, R., Purity of equivalued affine Springer fibers. Prépublication ; arXiv math.RT/0305141.

[9] Hales, T., The fundamental lemma for Sp(4). *Proc. Amer. Math. Soc.* **125** (1997), 301–308.

[10] Hitchin, N., Stable bundles and integrable connections. *Duke Math. J.* **54** (1987), 91–114.

[11] Kazhdan, D., et Lusztig, G., Fixed Point Varieties on Affine Flag Manifolds. *Israel J. Math.* **62** (1988), 129–168.

[12] Kottwitz, R., Stable trace formula : elliptic singular terms. *Math. Ann.* **275** (1986), 365–399.

[13] Kottwitz, R., Shimura varieties and λ-adic representations. Dans *Automorphic Forms, Shimura Varieties and L-functions* (Ann Arbor 1988), Perspect. Math. 10, Academic Press, Inc., Boston, MA, 1990, 161–209.

[14] Kottwitz, R., Calculation of some orbital integrals. Dans *The zeta functions of Picard modular surfaces* (ed. par. R. P. Langlands and D. Ramakrishnan), Université de Montréal, Centre de Recherches Mathématiques, Montreal, QC, 1992, 349–362.

[15] Kottwitz, R. Transfer factors for Lie algebras. *Represent. Theory* **3** (1999), 127–138.

[16] Kottwitz, R., Harmonic Analysis on semi-simple p-adic Lie algebras. Dans *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 553–562.

[17] Kottwitz, R., et Shelstad, D., Foundations of twisted endoscopy. *Astérisque* **225** (1999).

[18] Labesse, J.-P., et Langlands, R., L-indistinguishability for SL(2). *Canad. J. Math.* **31** (1979), 726–785.

[19] Langlands, R., *Les débuts d'une formule des traces stables*. Publications de l'Université Paris VII 13, 1983.

[20] Langlands, R., et Shelstad, D., On the definition of transfer factors. *Math. Ann.* **278** (1987), 219–271.

[21] Laumon, G., Fibres de Springer et jacobiennes compactifiées. Prépublication ; arXiv math.AG/0204109.

[22] Laumon, G., et Ngô, B.-C., Le lemme fondamental pour les groupes unitaires. Prépublication ; arXiv math.AG/0404454.

[23] Ngô, B.-C., Fibration de Hitchin et structure endoscopique de la formule des traces. Dans *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1213–1225.

[24] Rego, C. J., The Compactified Jacobian. *Ann. Sci. École Norm. Sup.* (4) **13** (1980), 211–223.

[25] Rogawski, J., *Automorphic representations of unitary groups in three variables.* Ann. of Math. Stud. 123, Princeton University Press, Princeton, NJ, 1990.

[26] Springer, T. A., A purity result for fixed point varieties in flag manifolds. *J. Fac. Sci. Univ. Tokyo* **31** (1984), 271–282.

[27] Waldspurger, J.-L., Sur les intégrales orbitales tordues pour les groupes linéaires : un lemme fondamental. *Canad. J. Math.* **43** (1991), 852–896.

[28] Waldspurger, J.-L., Comparaison d'intégrales orbitales pour des groupes p-adiques. Dans *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 807–816.

[29] Waldspurger, J.-L., Intégrales orbitales nilpotentes et endoscopie pour les groupes classiques non ramifiés. *Astérisque* **269** (2001).

[30] Waldspurger, J.-L., Endoscopie et changement de caractéristique. Prépublication.

[31] Weissauer, R., A special case of fundamental lemma. Prépublication.

Université Paris-Sud, Mathématiques, Bât. 425, 91405 Orsay Cedex, France

E-mail: gerard.laumon@math.u-psud.fr

Equidistribution, L-functions and ergodic theory: on some problems of Yu. Linnik

Philippe Michel and Akshay Venkatesh*

Abstract. An old question of Linnik asks about the equidistribution of integral points on a large sphere. This question proved to be very rich: it is intimately linked to modular forms, to subconvex estimates for L-functions, and to dynamics of torus actions on homogeneous spaces. Indeed, Linnik gave a partial answer using ergodic methods, and his question was completely answered by Duke using harmonic analysis and modular forms. We survey the context of these ideas and their developments over the last decades.

Mathematics Subject Classification (2000). Primary 11F66; Secondary: 11F67, 11M41.

Keywords. Automorphic L-functions, ergodic theory, equidistribution, subconvexity.

1. Linnik's problems

Given Q a homogeneous polynomial of degree m in n variables with integral coefficients, a classical problem in number theory is to understand the integral representations of an integer d by the polynomial Q, as $|d| \to +\infty$. Let $V_{Q,d}(\mathbb{Z}) = \{\boldsymbol{x} \in \mathbb{Z}^n, Q(\boldsymbol{x}) = d\}$ denote the set of such representations (possibly modulo some obvious symmetries). If $|V_{Q,d}(\mathbb{Z})| \to +\infty$ with d, it is natural to investigate the distribution of the discrete set $V_{Q,d}(\mathbb{Z})$ inside the affine variety "of level d"

$$V_{Q,d}(\mathbb{R}) = \{\boldsymbol{x} \in \mathbb{R}^n,\ Q(\boldsymbol{x}) = d\}.$$

In fact, one may rather consider the distribution, inside the variety of fixed level $V_{Q,\pm 1}(\mathbb{R})$, of the radial projection $|d|^{-1/m}.V_{Q,d}(\mathbb{Z})$ (here $\pm$ is the sign of d) and one would like to show that, as $|d| \to +\infty$, the set $|d|^{-1/m}.V_{Q,d}(\mathbb{Z})$ becomes equidistributed with respect to some natural measure $\mu_{Q,\pm 1}$ on $V_{Q,\pm 1}(\mathbb{R})$. Here, to take care of the case where $V_{Q,d}(\mathbb{Z})$ and $\mu_{Q,\pm 1}(V_{Q,\pm 1}(\mathbb{R}))$ are infinite, equidistribution w.r.t. $\mu_{Q,\pm 1}$ is defined by the following property: for any two sufficiently nice compact subsets $\Omega_1, \Omega_2 \subset V_{Q,\pm 1}(\mathbb{R})$ one has

$$\frac{\left||d|^{-1/m}.V_{Q,d}(\mathbb{Z}) \cap \Omega_1\right|}{\left||d|^{-1/m}.V_{Q,d}(\mathbb{Z}) \cap \Omega_2\right|} \longrightarrow \frac{\mu_{Q,\pm 1}(\Omega_1)}{\mu_{Q,\pm 1}(\Omega_2)} \quad \text{as } |d| \to +\infty. \tag{1.1}$$

*The research of the first author is partially supported by the Marie Curie RT Network "Arithmetic Algebraic Geometry" and by the "RAP" network of the Région Languedoc-Roussillon. The research of the second author is supported by a Clay Mathematics Research Fellowship and NSF grant DMS-0245606.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

The most general approach to this kind of problems is the circle method of Hardy–Littlewood. (Un)fortunately, that method is fundamentally limited to cases where the number of variables n is large compared with the degree m. To go further, one is led to make additional hypotheses on the varieties $V_{Q,d}$. It was anticipated by Linnik in the early 60's, and systematically suggested by Sarnak in the 90s [55], [56], [68], that for varieties which are homogeneous with respect to the action of some algebraic group $\boldsymbol{G}_{\mathbb{Q}}$, one should be able to take advantage of this action. Equidistribution problems on such homogeneous varieties are called (after Sarnak), equidistribution problems of Linnik's type.

By now, this expectation is largely confirmed by the resolution of wide classes of problems of Linnik's type ([10], [25], [30]–[32], [35], [57], [58], [65]); and the methods developed to deal with them rely heavily on powerful techniques from harmonic analysis (Langlands functoriality, quantitative equidistribution of Hecke points and approximations to the Ramanujan–Petersson conjecture) or from ergodic theory (especially Ratner's classification of measures invariant under unipotent subgroups), complemented by methods from number theory.

In this lecture we will not discuss that much the resolution of these important and general cases (for this we refer to [33], [72]); instead, we wish to focus on three, much older, examples of low dimension and degree ($m = 2$, $n = 3$) which were originally studied in the sixties by Linnik and his school. Our point in highlighting these examples is that the various methods developed to handle them are fairly different from the aforementioned ones which, in fact, may not apply or at least not directly.

The three problems correspond to taking Q to be a ternary quadratic form of signature $(3, 0)$ or $(1, 2)$. They are problems of Linnik's type with respect to the action of the orthogonal group $\boldsymbol{G} = \mathrm{SO}(Q)$ on $V_{Q,d}$.

The first problem is for the definite quadratic form $Q(A, B, C) = A^2 + B^2 + C^2$. For d an integer, $V_{Q,|d|}(\mathbb{Z})$ is the set of representations of $|d|$ as a sum of three squares

$$V_{Q,|d|}(\mathbb{Z}) = \{(a, b, c) \in \mathbb{Z}^3,\ a^2 + b^2 + c^2 = |d|\}$$

and $V_{Q,1}(\mathbb{R}) = S^2$ is the unit sphere. We denote by

$$\mathcal{G}_d = |d|^{-1/2}.V_{Q,|d|}(\mathbb{Z})$$

the radial projection of $V_{Q,|d|}(\mathbb{Z})$ on S^2:

Theorem 1 (Duke [17]). *For $d \to -\infty$, and $d \not\equiv 0, 1, 4 \pmod 8$ the set $\mathcal{G}_d$ is equidistributed on S^2 w.r.t. the Lebesgue measure μ_{S^2}.*

It will be useful to recall the "accidental" isomorphism of $\mathrm{SO}(Q)$ with $\boldsymbol{G} = \mathrm{PG}(\mathrm{B}^{(2,\infty)}) = \mathrm{B}^{\times}_{2,\infty}/Z(\mathrm{B}^{\times}_{2,\infty})$ where $\mathrm{B}^{(2,\infty)}$ is the algebra of the Hamilton quaternions. This arises from the identification of the quadratic space $(\mathbb{Q}^3, Q)$ with the trace-0 Hamilton quaternions endowed with the norm form $N(z) = z.\overline{z}$ via the map $(a, b, c) \to z = a.i + b.j + c.k$.

The second and third problems are relative to the indefinite quadratic form $Q(A, B, C) = B^2 - 4AC$, which is the discriminant of the binary quadratic forms $q_{A,B,C}(X, Y) = AX^2 + BXY + CY^2$. In that case, there is another "accidental" isomorphism of $\mathrm{SO}(Q)$ with PGL_2 via the map

$$(a, b, c) \to q_{a,b,c}(X, Y) = aX^2 + bXY + cY^2$$

which identifies $V_{Q,d}$ with the set $\mathcal{Q}_d$ of binary quadratic forms of discriminant d; PGL_2 acts on the latter by linear change of variables, twisted by inverse determinant. As $\mathrm{PGL}_2(\mathbb{Z})$ acts on $\mathcal{Q}_d(\mathbb{Z})$, one sees that, if $V_{Q,d}(\mathbb{Z}) = \mathcal{Q}_d(\mathbb{Z})$ is non empty (i.e. if $d \equiv 0, 1 \bmod 4$), it is infinite; so the proper way to define the equidistribution of $|d|^{-1/2}.V_{Q,d}(\mathbb{Z})$ inside $V_{Q,\pm 1}(\mathbb{R}) = \mathcal{Q}_{\pm 1}(\mathbb{R})$ is via (1.1). However, it is useful to formulate these problems in a slightly different (although equivalent) form which will be suitable for number theoretic applications. Let $\mathbb{H}^{\pm} = \mathbb{H}^+ \cup \mathbb{H}^- = \mathbb{C} - \mathbb{R} = \mathrm{PGL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R})$ denote the union of the upper and lower half-planes and $Y_0(1)$ denote the (non-compact) modular surface of full level i.e. $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathbb{H}^{\pm} \simeq \mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}^+$.

As is well known, the quotient $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathcal{Q}_d(\mathbb{Z})$ is finite, of cardinality some *class number* $h(d)$. For negative discriminants d, one associates to each $\mathrm{PSL}_2(\mathbb{Z})$-orbit $[q] \subset \mathcal{Q}_d(\mathbb{Z})$, the point $z_{[q]}$ in $Y_0(1)$ defined as the $\mathrm{PGL}_2(\mathbb{Z})$-orbit of the unique root of $q(X, 1)$ contained in $\mathbb{H}^+$. These points are called *Heegner points of discriminant*[1] d and we set

$$\mathcal{H}_d := \{z_{[q]},\ [q] \in \mathrm{PSL}_2(\mathbb{Z})\backslash\mathcal{Q}_d(\mathbb{Z})\} \subset Y_0(1).$$

An equivalent form to (1.1) for $Q(A, B, C) = B^2 - 4AC$ and $d \to -\infty$ is the following:

Theorem 2 (Duke [17]). *As $d \to -\infty$, $d \equiv 0, 1 \bmod 4$, the set $\mathcal{H}_d$ becomes equidistributed on $Y_0(1)$ w.r.t. the Poincaré measure $d\mu_P = \frac{3}{\pi}\frac{dxdy}{y^2}$.*

For positive discriminants d, one associates to each class of integral quadratic form $[q] \in \mathcal{Q}_d(\mathbb{Z})$ the positively oriented geodesic, $\gamma_{[q]}$, in $Y_0(1)$ which is the projection to $Y_0(1)$ of the geodesic line in $\mathbb{H}^+$ joining the two (real) roots of $q(X, 1)$. This is a closed geodesic – in fact, all closed geodesics on $Y_0(1)$ are of that form – whose length is essentially equal to the logarithm of the fundamental solution to Pell's equation $x^2 - dy^2 = 4$. We denote by

$$\Gamma_d := \{\gamma_{[q]},\ [q] \in \mathrm{PSL}_2(\mathbb{Z})\backslash\mathcal{Q}_d(\mathbb{Z})\}$$

the set of all geodesics of discriminant d.

Theorem 3 (Duke [17]). *As $d \to +\infty$, $d \equiv 0, 1 \bmod 4$, d not a perfect square, the set Γ_d becomes equidistributed on the unit tangent bundle of $Y_0(1)$, $S_1^*(Y_0(1))$, w.r.t. the Liouville measure $d\mu_L = \frac{3}{\pi}\frac{dxdy}{y^2}\frac{d\theta}{2\pi}$.*

[1] For simplicity, we will ignore non-primitivite forms.

These three problems (in their form (1.1)) were first proved by Linnik and by Skubenko by means of Linnik's *ergodic method*; we will return to this method in Section 6. The proof however is subject to an additional assumption which we call *Linnik's condition*, namely:

Theorem 4 (Linnik [54], [55], Skubenko [71]). *Let p be an arbitrary fixed prime, then the equidistribution statements of Theorems* 1, 2, 3 *hold for the subsequence of d such that p* is split *in the quadratic extension $K_d = \mathbb{Q}(\sqrt{d})$.*

We will see in Section 6 that Linnik's condition has a natural ergodic interpretation. It can be relaxed to the condition that for each d there is a prime $p = p(d) \leqslant |d|^{\frac{1}{10^{10}\log\log|d|}}$ which splits in $\mathbb{Q}(\sqrt{d})$. The latter condition is satisfied, for instance, by assuming that the L-functions of quadratic characters satisfying the Generalized Riemann Hypothesis (GRH). In particular, Linnik's condition (resp. the weaker one) is automatically fullfilled for subsequences of d such that K_d is a *fixed* quadratic field (resp. $\mathrm{disc}(K_d) = \exp\big(O\big(\frac{\log|d|}{\log\log|d|}\big)\big)$); however, in these cases, the proof of Theorems 1, 2, 3 is much simpler (see [11] for instance); so, as it is (from our perspective at least) the hardest case, we will limit ourselves to d's which are fundamental discriminants (i.e. $d = \mathrm{disc}(K_d)$).

Acknowledgements. The first author is scheduled to give a presentation based on this work in the ICM 2006. Since much of it is based on our joint work, we have decided to write this paper jointly. The results of Section 6 are all joint work with M. Einsiedler and E. Lindenstrauss and will also be discussed in their contribution to these proceedings [28].

It is our pleasure to thank Bill Duke, Henryk Iwaniec and Peter Sarnak for both their consistent encouragement and for many beautiful ideas which underlie the whole field. Peter Sarnak and Hee Oh carefully read an early draft and provided many helpful comments and corrections. We also would like to thank our collaborators Manfred Einsiedler and Elon Lindenstrauss, for patiently explaining ergodic ideas and methods to us.

2. Linnik's problems via harmonic analysis

Duke's unconditional solution of Linnik's problems is via harmonic analysis and in a sense, is very direct as it proceeds by verifying Weyl's equidistribution criterion. Let (X, μ) denote any of the probability spaces (S^2, μ_{S^2}), $(Y_0(1), \mu_P)$, $(S^1_*(Y_0(1)), \mu_L)$. For each case and for appropriate d, let μ_d denote the probability measure formed out of the respective sets $\mathcal{G}_d$, $\mathcal{H}_d$ or Γ_d: for instance for $X = S^2$,

$$\int_{S^2} \varphi \mu_d = \frac{1}{|\mathcal{G}_d|} \sum_{\substack{(a,b,c)\in\mathbb{Z}^3 \\ a^2+b^2+c^2=|d|}} \varphi\Big(\frac{a}{\sqrt{|d|}}, \frac{b}{\sqrt{|d|}}, \frac{c}{\sqrt{|d|}}\Big).$$

Showing that μ_d weak-$*$ converges to μ amounts to show that, for any φ ranging over a fixed orthogonal basis (made of continuous functions) of the L^2-space $L^2_0(X, \mu)$, the Weyl sum

$$W(\varphi, d) := \int_X \varphi \mu_d \quad \text{converges to 0 as } |d| \to +\infty. \tag{2.1}$$

In the context of Theorem 1 (resp. Theorem 2, resp. Theorem 3) such bases are taken to consist of non-constant harmonic polynomials (resp. Maass forms and Eisenstein series of weight 0, resp. Maass forms and Eisenstein series of non-negative, even, weight).

2.1. Duke's proof. The decay of the period integral $W(\varphi, d)$ is achieved by realizing it in terms of the d-th Fourier coefficient of a modular form of half-integral weight and level 4; this modular form – call it $\tilde{\varphi}$ – is obtained from φ through a theta correspondance.

In the case of Theorem 1, and when φ is a non-constant harmonic polynomial of degree r, this comes from the well known fact that the theta-series

$$\tilde{\varphi}(z) = \theta_\varphi(z) = \sum_{|d| \geqslant 1} \Big(\sum_{\substack{(a,b,c) \in \mathbb{Z}^3 \\ a^2+b^2+c^2 = |d|}} \varphi(a, b, c) \Big) e(|d|z)$$

is a modular form of weight $k = 3/2 + r$ for the modular group $\Gamma_0(4)$. This is a special case of a (theta) correspondance of Maass, which itself is now a special case of the theta correspondance for dual pairs; it associates to an automorphic form φ for an orthogonal group $\mathrm{SO}_{p,q}$ of signature (p, q), a Maass form $\tilde{\varphi}$ of weight $(q - p)/2$. Moreover, Maass provided a formula expressing the Fourier coefficients of $\tilde{\varphi}$ in terms of a certain integral of φ.

By the accidental isomorphisms recalled above, this provides a correspondance between automorphic forms either for $\mathrm{B}^\times_{2,\infty}$ or for PGL_2, and modular forms of half-integral weight. Under this correspondance, one has, for d a fundamental discriminant

$$W(\varphi, d) = c_{\varphi,d} \frac{\rho_{\tilde{\varphi}}(d)|d|^{-1/4}}{L(\chi_d, 1)} \tag{2.2}$$

where $c_{\varphi,d}$ is a constant depending on φ and mildly on d (i.e. $|d|^{-\varepsilon} \ll_{\varphi,\varepsilon} c_{\varphi,d} \ll_\varepsilon |d|^\varepsilon$ for any $\varepsilon > 0$), $\rho_{\tilde{\varphi}}(d)$ denotes the suitably normalized d-th fourier coefficient of $\tilde{\varphi}$ and χ_d is the quadratic character corresponding to K_d.

In particular, by Siegel's lower bound $L(\chi_d, 1) \gg_\varepsilon |d|^{-\varepsilon}$, (2.1) is a consequence of a bound of the form

$$\rho_{\tilde{\varphi}}(d) \ll |d|^{1/4-\delta} \tag{2.3}$$

for some absolute $\delta > 0$. The bound (2.3) is to be expected; indeed the half-integral weight analog of the Ramanujan–Petersson conjecture predicts that any $\delta < 1/4$ is admissible. This conjecture follows from the GRH, but, unlike its integral weight analogue, does not follow from the Weil conjectures.

The problem of bounding Fourier coefficient of modular forms can be approached through a Petersson–Kuznetzov type formula (due to Proskurin in the half-integral weight case): (un)fortunately the standard bound for the Salié sums occuring in the formula yield the above estimate only for $\delta < 0$. This "barricade" was eventually surmounted by Iwaniec (using an ingenious idea of averaging over the level, and obtaining the value $\delta = 1/28$, [41]) for $\tilde{\varphi}$ a holomorphic form of weight $\geqslant 5/2$ and by B. Duke for general forms by adapting Iwaniec's argument, and thus concluding the first fully unconditional proof of Theorems 1, 2, 3.

2.2. Equidistribution and subconvex bounds for L-functions. Shortly after Duke's proof, another approach emerged which turned out to be very fruitful, namely the connection between the decay of Weyl's sums (2.1) and the *subconvexity problem* for automorphic L-function (see Section 3).

2.2.1. Weyl's sums as period integrals: Waldspurger type formulae. It goes back to Gauss that the set of classes of quadratic forms $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{Q}_d(\mathbb{Z})$ has the structure of a finite commutative group (the class group) $\mathrm{Cl}(d)$. In particular for the second problem ($d < 0$), $\mathcal{H}_d$ is a homogeneous space under the action of $\mathrm{Cl}(d)$ and the Weyl sums can be seen as period integrals for this action:

$$W(\varphi, d) = \int_{\mathrm{Cl}(d)} \varphi(\sigma.z_d)\, d\mu_{\mathrm{Haar}}(\sigma).$$

In a similar way, the Weyl's sums over $\mathcal{G}_d$ and Γ_d can be realized as orbital integrals for the action of some class group. The connection between such orbital integrals and L-functions follows from a formula basically due to Waldspurger. To describe it in greater detail it is useful and convenient to switch an adelic description of the Weyl's sums.

Let us recall recall that in the context of Theorem 1 with $Q(A, B, C) = A^2 + B^2 + C^2$ (resp. Theorems 2 and 3, with $Q(A, B, C) = B^2 - 4AC$) a solution $Q(a, b, c) = d$ gives rise to an embedding of the quadratic $\mathbb{Q}$-algebra K_d into the $\mathbb{Q}$-algebra $\mathrm{B}^{(2,\infty)}$ (resp. $M_{2,\mathbb{Q}}$) by sending $\sqrt{d}$ to $a.i + b.j + c.k$ (resp. $\left(\begin{smallmatrix} b & -2a \\ 2c & -b \end{smallmatrix}\right)$). This yields an embedding of $\mathbb{Q}$-algebraic groups, $\boldsymbol{T}_d := \mathrm{res}_{K/Q}\mathbb{G}_m/\mathbb{G}_m \hookrightarrow \boldsymbol{G}$, where $\boldsymbol{G} = \mathrm{PG}(\mathrm{B}^{(2,\infty)})$ (resp. $= \mathrm{PGL}_2$).

Let $\mathrm{K}_{f,\max}$ be a maximal compact subgroup of $\boldsymbol{G}(\boldsymbol{A}_f)$ in all three cases. In the context of Theorem 1 (resp. Theorem 2, resp. Theorem 3) take $\mathrm{K}_\infty = \boldsymbol{T}_d(\mathbb{R}) \cong \mathrm{SO}_2 \subset \boldsymbol{G}$ (resp. $\mathrm{K}_\infty = \boldsymbol{T}_d(\mathbb{R})$, resp. $\mathrm{K}_\infty = \{1\}$) and set $\mathrm{K} = \mathrm{K}_{f,\max}\mathrm{K}_\infty$; the quotient $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A}_\mathbb{Q})/\mathrm{K}$ then equals a quotient of S^2 by a finite group of rotations (resp. $Y_0(1)$, resp. the unit tangent bundle of $Y_0(1)$).

It transpires, with these identifications, the subsets

$$\mathcal{G}_d \subset S^2, \quad \mathcal{H}_d \subset Y_0(1), \quad \Gamma_d \subset S^1_*(Y_0(1))$$

may be uniformly described, after choosing a solution z_d, as a compact orbit of the adelic torus $\boldsymbol{T}_d$:

$$\boldsymbol{T}_d(\mathbb{Q})\backslash z_d.\boldsymbol{T}_d(\boldsymbol{A}_{\mathbb{Q}})/\mathrm{K}_{\boldsymbol{T}_d} \subset \boldsymbol{G}(\mathbb{Q})\backslash \boldsymbol{G}(\boldsymbol{A}_{\mathbb{Q}})/\mathrm{K}$$

where $\mathrm{K}_{\boldsymbol{T}_d} := \boldsymbol{T}_d(\boldsymbol{A}_{\mathbb{Q}}) \cap \mathrm{K}$. In this notation the Weyl sum is given as a *toric integral*

$$W(\varphi, d) = \int_{\boldsymbol{T}_d(\mathbb{Q})\backslash \boldsymbol{T}_d(\boldsymbol{A}_{\mathbb{Q}})/K_{\boldsymbol{T}_d}} \varphi(z_d.t)\, dt \tag{2.4}$$

when dt is the Haar measure on the toric quotient. A superficial advantage of this notation is that it allows for a uniform presentation of many equidistribution problems for "cycles" associated with quadratic orders in locally symmetric spaces associated to quaternion algebras. Indeed, as we shall see below, one can consider the above equidistribution problems while changing

- the group $\boldsymbol{G}$ to $\boldsymbol{G} = \mathrm{B}^\times/Z(\mathrm{B}^\times)$ for B any quaternion algebra over $\mathbb{Q}$;
- the compact $\mathrm{K}_{f,\max}$ to a compact subgroup $\mathrm{K}'_f \subset \mathrm{K}_{f,\max}$ (i.e. changing the level structure)
- the subgroup $\mathrm{K}_{\boldsymbol{T}_d}$ to a subgroup $\mathrm{K}'_{\boldsymbol{T}_d}$ (i.e. considering cycles associated to suborders $\mathcal{O}$ of the maximal order $\mathcal{O}_d$).
- the base field $\mathbb{Q}$ to a fixed totally real number field F.

When φ is a *new cuspform* (the L^2 normalized *new vector* in some automorphic representation π), Waldspurger's formula [76] relates $|W(\varphi, d)|^2$ (and correspondingly the square of the d-th Fourier coefficient $|\rho_{\tilde{\varphi}}(d)|^2$) to the central value of an automorphic L-function. In its original form, the formula was given up to some non-zero proportionality constant; as we are interested in the size $W(\varphi, d)$ a more precise expression is needed. Thanks to the work of many people ([12], [36], [37], [47], [50], [66], [78], [80], [81]) notably Gross, Zagier and Zhang such an expression is by now available in considerable generality. Under suitable hypotheses (which in the present cases are satisfied), it has the following form

$$|W(\varphi, d)|^2 = c_{\varphi,d} \frac{L(\pi, 1/2)L(\pi \times \chi_d, 1/2)}{L(\chi_d, 1)^2\sqrt{|d|}} \tag{2.5}$$

where π' is a GL_2-automorphic representation corresponding to π by the Jacquet–Langlands correspondance and $c_{\varphi,d} > 0$ is a constant which depends mildy on d.

The Waldspurger formula (2.5) is more powerful than (2.2) as it may be extended to a formula for more general toric integrals. Indeed, let χ be a character of the torus $\boldsymbol{T}_d(\mathbb{Q})\backslash \boldsymbol{T}_d(\boldsymbol{A}_{\mathbb{Q}}) = K_d^\times \boldsymbol{A}_{\mathbb{Q}}^\times \backslash \boldsymbol{A}_{K_d}^\times$ trivial on $\mathrm{K}_{\boldsymbol{T}_d}$. Under suitable compatibility assumptions between χ and φ and possibly under additional coprimality assumptions between the conductors of of π, χ, the relation (2.5) generalizes to

$$|W_\chi(\varphi, d)|^2 = c_{\varphi,d_\chi,\chi_\infty} \frac{L(\pi \times \pi_\chi, 1/2)}{L(\chi_d, 1)^2\sqrt{|d_\chi|}} \tag{2.6}$$

where $W_\chi(\varphi, d)$ is a *twisted toric integral* of the form

$$W_\chi(\varphi, d_\chi) = \int_{\boldsymbol{T}_d(\mathbb{Q})\backslash \boldsymbol{T}_d(\boldsymbol{A}_\mathbb{Q})} \chi(t)\varphi(z_{d_\chi}.t)\, dt,$$

π_χ is the GL_2-automorphic representation (of conductor d_χ) corresponding to χ by quadratic automorphic induction and $L(\pi \times \pi_\chi, s)$ is the Rankin–Selberg L-function of the pair (π, π_χ).

2.3. Subconvexity and (sparse) equidistribution. We see, from formula (2.5) and Siegel's lower bound that (2.1) follows from the bound

$$L(\pi \times \chi_d, 1/2) \ll_\pi |d|^{1/2-\delta} \tag{2.7}$$

for some absolute $\delta > 0$; subject to this bound, one obtains another proof of Linnik's equidistribution problems. More generally, we see from (2.6) that the twisted Weyl sums are decaying, i.e.

$$W_\chi(\varphi, d_\chi) \to 0 \quad \text{for } d_\chi \to +\infty, \tag{2.8}$$

as soon as

$$L(\pi \times \pi_\chi, 1/2) \ll |d_\chi|^{1/2-\delta}. \tag{2.9}$$

Both (2.7) and (2.9) are special cases of subconvex bounds for central values of automorphic L-functions and have been proven (see below).

One should note that the decay of the twisted toric integral is useful if one needs to perform *harmonic analysis along* the toric orbit $\boldsymbol{T}_d(\mathbb{Q})\backslash z_d.\boldsymbol{T}_d(\boldsymbol{A}_\mathbb{Q})/\mathrm{K}_{\boldsymbol{T}_d}$: this is particular the case when one needs equidistribution only for a stricly smaller suborbit of the full orbit, a problem we call a *sparse equidistribution problem*.

For instance one has:

Theorem 5 ([39]). *There is an absolute constant $0 < \eta < 1$ such that: for each fundamental discriminant $d < 0$, choose $z_{0,d} \in \mathcal{H}_d$ a Heegner point and choose G_d a subgroup of $\mathrm{Cl}(d)$ of size $|G_d| \geqslant |\mathrm{Cl}(d)|^\eta$ then the sequence of suborbits*

$$\mathcal{H}'_d := G_d.z_{0,d} = \{\sigma.z_{0,d},\ \sigma \in G_d\}$$

is equidistributed on $Y_0(1)$ w.r.t. μ_P.

One has also similar sparse equidistribution results for sufficiently large suborbits of $\mathcal{G}_d$ on the sphere and for suffciently large geodesic segments of Γ_d [60], [66]. Note however that the present method has fundamental limitations as one cannot take η too close to 0: even under the GRH, one would prove equidistribution only for $\eta > 1/2$. Nevertheless we would like to formulate the following

Conjecture 1 (Equidistribution of subgroups). Fix any $\eta > 0$ and for each fundamental discriminant $d < 0$, choose $z_{0,d} \in \mathcal{H}_d$ a Heegner point and choose G_d a subgroup of $\mathrm{Cl}(d)$ of size $|G_d| \geqslant |d|^\eta$. Then as $|d| \to +\infty$, the sequence of suborbits

$$\mathcal{H}'_d := G_d.z_{0,d} = \{\sigma.z_{0,d},\ \sigma \in G_d\}$$

is equidistributed on $Y_0(1)$ w.r.t. μ_P.

This conjecture is certainly difficult in general; however, we expect that, by ergodic methods like the ones described in Section 6, significant progress might be made, at least for subgroups G_d that satisfy suitable versions of Linnik's condition for some fixed prime p.

2.4. Equidistribution and non-vanishing of L-functions. Before continuing with the subconvexity problem, we would like to point out another interesting application. It combines subconvexity, equidistribution and the period relation (2.5) and applies them to the non-vanishing of L-functions.

Consider, for simplicity, the context of Theorem 2 (see also [62]): let φ be a Maass–Hecke eigenform of weight 0 and π be its associated automorphic representation. If one averages (2.5) over the characters of $\mathrm{Cl}(d)$, one obtains by orthogonality (here the constants $c_{\varphi,d_\chi,\chi_\infty}$ are equal to an absolute constant $c > 0$)

$$c\frac{\sqrt{d}}{|\mathrm{Cl}(d)|^2}\sum_{\chi\in\widehat{\mathrm{Cl}}(d)} L(\pi\times\pi_\chi, 1/2) = \int_{Y_0(1)} |\varphi|^2.\mu_d$$

and since by Theorem 2

$$\int_{Y_0(1)} |\varphi|^2.\mu_d \to \int_{Y_0(1)} |\varphi(z)|^2 d\mu_P(z) > 0 \quad \text{as } d \to -\infty$$

this shows that for some χ the central value $L(\pi\times\pi_\chi, 1/2)$ does not vanish. Moreover, by the subconvex bound (2.9), one obtains a quantitative form of non-vanishing

$$\left|\{\chi \in \widehat{\mathrm{Cl}}(d),\ L(\pi\times\pi_\chi, 1/2) \neq 0\}\right| \gg |d|^\eta \tag{2.10}$$

for some absolute $\eta > 0$.

Remark 2.1. When π corresponds to an Eisenstein series, stronger results where obtained before by Duke–Friedlander–Iwaniec [24] and Blomer [5]; although this it appears in a somewhat disguised (and more elaborate) form, the basic principle underlying the proof is the same.

By considering equidistribution relative to definite quaternion algebras, one can obtain similar non-vanishing results for central values $L(\pi\times\pi_\chi, 1/2)$ where π_∞ is in the discrete series and the sign of the functional equation of $L(\pi\times\pi_\chi, s)$ is $+1$. In particular when $\pi = \pi_E$ is the automorphic representation associated to an elliptic

curve $E/\mathbb{Q}$, such estimates provide a lower bound for the size of the "rank-0" part of the group $E(H_K)$ of points of E which are rational over the Hilbert class field of K as $d \to -\infty$.

An interesting problem is to adress the case where the sign of the functional equation is -1. In this case, $L(\pi \times \pi_\chi, 1/2) = 0$ and one considers instead the question of non-vanishing of the first derivative $L'(\pi \times \pi_\chi, 1/2)$. At least when π_∞ is in the holomorphic discrete series and π has trivial central character, the Gross–Zagier formula (and its extensions by Zhang) interprets $L'(\pi \times \pi_\chi, 1/2)$ as the "height" of some Heegner cycle above some modular (or Shimura) curve. This is not quite a period integral; however the height decomposes as a sum of local heights indexed by the places v of $\mathbb{Q}$. These local heights are either simple or can be interpreted as periods integrals over quadratic cycles associated with K which live over appropriate adelic quotients $\boldsymbol{G}^{(v)}(\mathbb{Q})\backslash\boldsymbol{G}^{(v)}(\boldsymbol{A})/\mathrm{K}_v$ where $\boldsymbol{G}^{(v)}$ is associated to a quaternion algebra $\mathrm{B}^{(v)}$ ramified at v.

It seem then plausible that one can compute the asymptotic of the average $\sum_\chi L'(\pi \times \pi_\chi, 1/2)$ by using the equidistribution property of quadratic cycles on these infinitely many quotients. One consequence of this would then be, for compatible E and K, a lower bound for the rank of $E(H_K)$:

$$\mathrm{rank}_{\mathbb{Z}}\, E(H_K) \gg |d|^\eta$$

for some $\eta > 0$ as $d \to -\infty$.

Remark 2.2. A few years ago, Vatsal and Cornut [16], [73], [74] used period relations and equidistribution in a similar way to obtain somewhat stronger non-vanishing results for Rankin–Selberg L-functions but associated to anti-cyclotomic[2] characters of a *fixed* imaginary quadratic field. Note that one of their main ingredient to obtain equidistribution came for ergodic theory and precisely from Ratner's theory.

3. The subconvexity problem

Although the subconvexity problem is a venerable topic in number theory – its study begins with Weyl's estimate $|\zeta(1/2 + it)| \ll_\varepsilon t^{1/6+\varepsilon}$ – there has been a renaissance of interest in it recently. This owes largely to the observation that a resolution of the subconvexity problem for automorphic L-functions on GL has many striking applications, as we have just seen to Linnik's equidistribution problems or to "Arithmetic Quantum Chaos." We refer to [44] for a discussion of all these questions in the broader context of the analytic theory of automorphic L-functions.

Let $\Pi = \Pi_\infty \otimes \bigotimes_p{}' \Pi_p$ some reasonable "automorphic object": by automorphic object we mean, for instance an automorphic representation or more generally an admissible representation constructed out of automorphic representations via the

[2]The case of cyclotomic characters was carried out even earlier by Rohrlich, by more direct methods.

formalism of L-groups (for instance the Rankin–Selberg convolution $\pi_1 \times \pi_2$ of two automorphic representations on some linear groups). To Π, one can usually associate a collection of local L-factors

$$L(\Pi_p, s) = \prod_{i=1}^{d} \left(1 - \frac{\alpha_{\Pi,i}(p)}{p^s}\right)^{-1}, \quad p \text{ prime}, \ L(\Pi_\infty, s) = \prod_{i=1}^{d} \Gamma_{\mathbb{R}}(s - \mu_{\Pi,i})$$

where $\Gamma_{\mathbb{R}}(s) = \pi^{-s/2}\Gamma(s/2)$ and $\{\alpha_{\Pi,i}(p)\}$, $\{\mu_{\Pi,i}\}$ are called the local numerical parameters of Π at p and at infinity; from these local datas one forms a global L-function

$$L(\Pi, s) = \sum_{n \geqslant 1} \frac{\lambda_\Pi(n)}{n^s} = \prod_p L(\Pi_p, s).$$

In favourable cases, one can show that $L(\Pi, s)$ has analytic continuation to $\mathbb{C}$ and satisfies a functional equation which we normalize into the form

$$q_\Pi^{s/2} L(\Pi_\infty, s) L(\Pi, s) = w_\Pi q_\Pi^{(1-s)/2} \overline{L(\Pi_\infty, 1 - \overline{s}) L(\Pi, 1 - \overline{s})},$$

where $|w_\Pi| = 1$ and $q_\Pi > 0$ is an integer called the conductor of Π. We recall (after Iwaniec–Sarnak [44]) that *the analytic conductor* of Π is the function of the complex variable s given by

$$C(\Pi, s) = q_\Pi \prod_{i=1}^{d} |s - \mu_{\Pi,i}|.$$

It is expected, and known in many cases, that the following *convexity bound* for the values of $L(\Pi, s)$ holds on the critical line $\Re es = 1/2$: for any $\varepsilon > 0$, one has

$$L(\Pi, s) \ll_{\varepsilon,d} C(\Pi, s)^{1/4+\varepsilon}.$$

This is known, in particular, when Π is an automorphic cuspidal representation of $\mathrm{GL}(n)$ over any number field, [63]. The Lindelöf conjecture, which is a consequence of the GRH, asserts that in fact $L(\Pi, s) \ll_{\varepsilon,d} C(\Pi, s)^{\varepsilon}$. In many applications, however, it is sufficient to improve the convexity bound.

The subconvexity problem consists in improving the exponent $1/4$ to $1/4 - \delta$ for some positive absolute δ. In fact, for most applications it is sufficient to improve that exponent only with respect to one of the three type of parameters s, q_Π or $\prod_{i=1}^{d}(1 + |\mu_{\Pi,i}|)$; these variants are called the s-aspect, the q-aspect (or *level*-aspect) and the ∞-aspect (or *eigenvalue*-aspect) respectively. See [34], [44] for an introduction to the subconvexity problem in this generality. During the last decade, there has been considerable progress on the subconvexity problem for L-functions associated to GL_1 and GL_2 automorphic forms. In this lecture, we mainly discuss the recent progress made on the q-aspect, although the other aspects are very interesting, both for applications and for conceptual reasons (see [6], [42], [46], [70]). In the level aspect, one has

Theorem 6. *Let F be a fixed number field and π_2 be a* fixed *cuspidal automorphic representation of $\mathrm{GL}_2(\boldsymbol{A}_F)$. Let χ_1, π_1 denote respectively a $\mathrm{GL}_1(\boldsymbol{A}_F)$-automorphic representation (i.e. a Grössencharacter), a $\mathrm{GL}_2(\boldsymbol{A}_F)$-automorphic representation and let $\mathfrak{q}_1$ denote either the conductor of χ_1 or π_1 and $q_1 = N_{F/\mathbb{Q}}(\mathfrak{q}_1)$. There exists an absolute constant $\delta > 0$ such that for $\Re es = 1/2$ one has*

$$L(\chi_1, s) \ll_s q_1^{1/4-\delta}, \tag{3.1}$$

$$L(\chi_1 \times \pi_2, s) \ll_{s,\pi_2,\chi_{1,\infty}} q_1^{1/2-\delta}, \tag{3.2}$$

$$L(\pi_1, s) \ll_{s,\pi_{1,\infty}} q_1^{1/4-\delta}, \tag{3.3}$$

$$L(\pi_1 \times \pi_2, s) \ll_{s,\pi_2,\pi_{1,\infty}} q_1^{1/2-\delta}. \tag{3.4}$$

Thus the subconvexity problem is solved in the q_1-aspect for all these L-functions.

• For $F = \mathbb{Q}$, the bound for Dirichlet L-functions (3.1) is due to Burgess (see also [15]). The bound for twisted L-function (3.2) is basically due to Duke–Friedlander–Iwaniec [19] (see also [7], [9] for the general bound over $\mathbb{Q}$ with a good subconvex exponent). The bound (3.3) is mainly to a series of works by Duke–Friedlander–Iwaniec: [20] for π_1 with trivial central character and [21], [22], [23] for the much harder case of a central character of conductor q_1; it has been recently completed for π_1 with arbitrary central character by Blomer, Harcos and the first author in [8]. The bound for Rankin–Selberg L-functions (3.4) for π having trivial central character is due to Kowalski, the first author and Vanderkam ([52]) by generalizing the methods of [20] and to Harcos and the first author for π_1 with an arbitrary central character [39], [60].

• In the case of a number field of higher degree, the first general subconvex result is due to Cogdell–Piatetski-Shapiro–Sarnak [13]: it consists of (3.2) when F is a totally real field and $\pi_{2,\infty}$ is in the holomorphic discrete series (i.e. corresponds to a holomorphic Hilbert modular form). Recently, the second author developed a new method and established, amongst other things, the bounds (3.1), (3.2), (3.3) and (3.4) for F an arbitrary number field, π_2 fixed but arbitrary and π_1 with a trivial central character [75]. Eventually the authors combined their respective methods from [60] and [75] to obtain (3.3) and (3.4) for π_1 with an arbitrary central character.

3.1. Amplification and the shifted convolution problem. Arguably, the most successful approach to subconvexity in the q-aspect is via the method of moments or more precisely via its variant, the *amplification method*. For the sake of completness we briefly recall the mechanism and refer to [34] and [43] for the philosophy underlying this method.

Given Π_1 and a (well choosen) family of automorphic objects $\mathcal{F} = \{\Pi\}$ containing Π_1, the amplification method builds on the possibility to obtain a bound for the amplified k-th moment of the $\{L(\Pi, s),\ \Pi \in \mathcal{F}\}$, $\Re es = 1/2$, of the form

$$\sum_{\Pi\in\mathcal{F}} |L(\Pi, s)|^k \Big| \sum_{\ell\leqslant L} \lambda_\Pi(\ell) a_\ell \Big|^2 \ll_\varepsilon |\mathcal{F}|^{1+\varepsilon} \sum_{\ell\leqslant L} |a_\ell|^2 \tag{3.5}$$

for any $\varepsilon > 0$, where the $(a_\ell)_{\ell \leqslant L}$ are *a priori* arbitrary complex coefficients and where L is some positive power of $|\mathcal{F}|$. Such a bound is expected if L is sufficiently small compared with $|\mathcal{F}|$, since the individual bound $|L(\Pi, s)|^k \ll_\varepsilon |\mathcal{F}|^\varepsilon$ would follow from the GRH and the estimate

$$\sum_{\Pi \in \mathcal{F}} \Big| \sum_{\ell \leqslant L} \lambda_\Pi(\ell) a_\ell \Big|^2 = |\mathcal{F}|(1 + o(1)) \sum_{\ell \leqslant L} |a_\ell|^2$$

should be a manisfestation of the *quasi-orthogonlity* of the $\{(\lambda_\Pi(\ell))_{\ell \leqslant L}\}_{\Pi \in \mathcal{F}}$ which is a frequently recurring theme in harmonic analysis. Assuming (3.5), one deduces a subconvex bound for $|L(\Pi_1, s)|^k$ by restricting (3.5) to one term and by choosing the coefficients $a_\ell = a_\ell(\Pi_1)$ appropriately.

All the subconvex bounds presented in Theorem 6 can be obtained by considering for $L(\Pi, s)$ an L-function of *Rankin–Selberg* type, i.e. an L-function either of the form $L(\chi_1 \times \pi_2, s)$ or of the form $L(\pi_1 \times \pi_2, s)$ with π_2 a fixed (not necessarily cuspidal) GL_2-automorphic representation. The families $\mathcal{F}$ considered are then essentially of the form $\{\chi \times \pi_2,\ q_\chi = q_1\}$ or $\{\pi \times \pi_2,\ q_\pi = q_1,\ \omega_\pi = \omega_{\pi_1}\}$ and the bound (3.5) is achieved for the second moment ($k = 2$). To analyze effectively the left-hand side of (3.5) one needs a manageable expression for $L(\Pi, s)$ for s on the critical line. The traditional method to do so is to apply an *approximate functional equation* technique which expresses $L(\Pi, s)$ essentially as a partial sum of the form

$$\Sigma(\Pi) := \sum_{n \geqslant 1} \frac{\lambda_\Pi(n)}{n^s} W\left(\frac{n}{\sqrt{q_\Pi}}\right)$$

with W a rapidly decreasing function (which depends on s and on Π_∞). In the context of Theorem 6, the second amplified moment (3.5) are then computed and transformed by spectral methods. These involve, in particular, the orthogonality relations for characters and the Kuznetsov–Petersson formula. These computations reduce the subconvex estimates to the problem of estimating non-trivially sums of the form

$$\Sigma_\pm(\varphi_2, \ell_1, \ell_2, h) := \sum_{\ell_1 m \pm \ell_2 n = h} \overline{\rho_{\varphi_2}(m)} \rho_{\varphi_2}(n) \mathcal{W}\left(\frac{m}{q}, \frac{n}{q}\right), \tag{3.6}$$

the trivial bound being $\ll_{\varphi_2} q^{1+o(1)}$; here $h = O(q)$ is a non-zero integer, $\rho_{\varphi_2}(n)$ denotes the n-th Fourier coefficient of some automorphic form φ_2 in the representation space of π_2, $\mathcal{W}(x, y)$ is a rapidly decreasing function and $\ell_1, \ell_2 \leqslant L$ are the parameters occuring as indices of the amplifier $(a_\ell)_{\ell \leqslant L}$. These sums are classical in analytic number theory and are called *shifted convolution sums*; the problem of estimating them non-trivially for various ranges of h, m, n is called a *shifted convolution problem*.[3]

[3]Historically, the shifted convolution problem already occured in the work of Kloosterman on the number of representations of an integer n by the quadratic form $a_1.x^2 + a_2.y^2 + a_3.z^2 + a_4.t^2$, and also in Ingham's work on the additive divisor problem. In Kloosterman's case φ_2 is a theta-series of weight 1, whereas in Ingham's case φ_2 is the standard non-holomorphic Eisenstein series.

3.2. Shifted convolutions via the circle method. In order to solve a shifted convolution problem, one needs an analytically manageable expression of the linear constraint $\ell_1 m \pm \ell_2 n = h$; one is to suitably decompose the integral

$$\delta_{\ell_1 m \pm \ell_2 n - h = 0} = \int_{\mathbb{R}/\mathbb{Z}} \exp(2\pi i (\ell_1 m \pm \ell_2 n - h)\alpha)\, d\alpha,$$

and there are several methods to achieve this; the first possibility in this context was Kloosterman's refinement of the circle method; other possibilities are the Δ-symbol method, used in [19] and [20] to prove some cases of (3.2) and (3.3) or Jutila's method of overleaping intervals which is particularly flexible [38], [45]. These methods provide an expression of the above integral into weighted sums of Ramanujan type sums of the form

$$\sum_{\substack{a \bmod c \\ (a,c)=1}} e\Big(\frac{(\ell_1 m \pm \ell_2 n - h).a}{c}\Big)$$

for c ranging over relatively small moduli. Such decomposition makes it possible to essentially "separate" the variable m from n and to reduce $\Sigma(\varphi_2, \ell_1, \ell_2, h)$ to sums over moduli c on additively twisted sums of Fourier coefficients

$$\sum_c \cdots \sum_{\substack{a \bmod c \\ (a,c)=1}} e\Big(\frac{-ha}{c}\Big)\Big(\sum_m \overline{\rho_{\varphi_2}(m)} e\Big(\frac{\ell_1 m a}{c}\Big) W\Big(\frac{m}{q}\Big)\Big)$$
$$\Big(\sum_n \rho_{\varphi_2}(n) e\Big(\pm\frac{\ell_2 n a}{c}\Big) W\Big(\frac{n}{q}\Big)\Big).$$

The independent m- and n-sums are then transformed via the Voronoï summation formula with the effect of replacing the test functions $W(\frac{\cdot}{q})$ by some Bessel transform and the additive shift $e\big(\pm\frac{\ell_2 a.}{c}\big)$ by $e\big(-\pm\frac{\bar{\ell}_2 \bar{a}.}{c}\big)$ where $\bar{a}$ denotes the multiplicative inverse of a mod c. After these transformations and after averaging over a mod c the sum $\Sigma_\pm(\varphi_2, \ell_1, \ell_2, h)$ takes essentially the following form (possibly up to a main term which occurs if φ_2 is an Eisenstein series [18])

$$MT_\pm(\varphi_2, \ell_1, \ell_2, h) \tag{3.7}$$
$$+ \sum_{c \equiv 0(\ell_1 \ell_2 q \pi_0)} \sum_{h'} \Big(\sum_{\mp \ell_1 n - \ell_2 m = h'} \alpha_m \overline{\rho_{\varphi_2}(m)} \beta_n \rho_{\varphi_2}(n) \Big) \mathrm{Kl}(-h, h'; c) \mathcal{V}(h, h'; c)$$

where $MT_\pm(\varphi_2, \ell_1, \ell_2, h)$ is non-zero only if φ_2 is an *Eisenstein series* (in which case it is a *main term* of size $\approx_{\ell_1, \ell_2, \varphi_2} q^{1+o(1)}$), α_m, β_n are smooth coefficients, $\mathrm{Kl}(-h, h'; c)$ is a Kloosterman sum and $\mathcal{V}$ is a smooth function. Eventually, Weil's bound for Kloosterman sums

$$\mathrm{Kl}(-h, h'; c) \ll (h, h', c)^{1/2} c^{1/2 + o(1))}$$

gives the formula

$$\Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h) = MT_{\pm}(\varphi_2, \ell_1, \ell_2, h) + O_{\varphi_2,\varepsilon}\big((\ell_1\ell_2)^A q^{3/4+\varepsilon}\big) \tag{3.8}$$

for some absolute constant A. Finally, from (3.8) one can deduce (3.2), (3.3), (3.4) when π has trivial central character although the derivation may be quite delicate if φ_2 is an Eisenstein series (cf. [20] and see also [51]).

3.3. Shifted convolutions and spectral theory. In [70], Sarnak, inspired by ideas of Selberg, developed a purely spectral approach to the shifted convolution sums (3.6) (previously some special cases have been treated by others, for instance by A. Good). This method, which at present has been entirely worked out when φ_2 is a classical holomorphic cuspform (say of weight $k \geqslant 2$ and level q_2), is based on the analytic properties of the Dirichlet series

$$D(\varphi_2, \ell_1, \ell_2, h, s) = \sum_{\substack{m,n\geqslant 1\\ \ell_1 m-\ell_2 n=h}} \frac{\overline{\rho_{\varphi_2}(m)}\rho_{\varphi_2}(n)}{(\ell_1 m+\ell_2 n)^s}\left(\frac{\sqrt{\ell_1\ell_2 mn}}{\ell_1 m+\ell_2 n}\right)^{k-1}.$$

Note that for $h = 0$ this series is essentially a Rankin–Selberg L-function. As in the Rankin–Selberg case, the analytic properties of D follows from an appropriate integral representation in the form of a triple product integral; however, for $h \neq 0$ one needs to replace the Eisenstein series by a Poincaré series. Precisely, one has $D(s) = (2\pi)^{s+k-1}(\ell_1\ell_2)^{1/2}\Gamma^{-1}(s+k-1)I(s)$ with

$$\begin{aligned} I(\varphi_2, \ell_1, \ell_2, h, s) &:= \langle (\ell_1 y)^{k/2}\varphi_2(\ell_1 z).(\ell_2 y)^{k/2}\overline{\varphi_2}(\ell_2 z),\, P_h(z,s)\rangle \\ &= \int_{\Gamma_0(q_2\ell_1\ell_2)\backslash\mathbb{H}} (\ell_1 y)^{k/2}\overline{\varphi_2}(\ell_1 z).(\ell_2 y)^{k/2}\varphi_2(\ell_2 z)P_h(z,s)\,\frac{dxdy}{y^2} \end{aligned}$$

where $P_h(z,s)$ is a non-holomorphic Poincaré series of weight 0. The analytic continuation for D follows from that of $P_h(.,s)$; in particular, from its spectral expansion one deduce that the latter is absolutely convergent for $\Re es > 1$ and has holomorphic continuation in the half-plane $\Re es > 1/2 + \theta$ where θ measures the quality of available results towards the Ramanujan–Petersson conjecture:

Hypothesis H_θ. For any cuspidal automorphic form π on $\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\boldsymbol{A}_{\mathbb{Q}})$ with local Hecke parameters $\{\alpha_{\pi,i}(p),\ i = 1, 2\}$ for $p < \infty$ and $\{\mu_{\pi,i},\ i = 1, 2\}$ one has the bounds

$$\begin{aligned} |\alpha_{\pi,i}(p)| &\leqslant p^{\theta}, \quad i = 1, 2,\\ |\Re e\mu_{\pi,i}| &\leqslant \theta, \qquad i = 1, 2, \end{aligned}$$

provided π_p, π_∞ are unramified, respectively.

Remark 3.1. Hypothesis H_θ is known for $\theta > 3/26$ thanks to the works of Kim and Shahidi [48], [49].

A bound for $D(s)$ in a non-trivial domain is deduced from the spectral expansion of the inner product $I(s)$ over an suitable orthonormal basis of Maass forms, $\{\psi\}$ say, and of Eisenstein series of weight 0 and level $\ell_1\ell_2 q_0$: one has

$$\sum_{\psi}\langle((\ell_1 y)^{k/2}\varphi_2(\ell_1 z).(\ell_2 y)^{k/2}\overline{\varphi_2(\ell_2 z)},\ \psi\rangle\langle\psi, P_h(z,s)\rangle + \text{Eisenstein spectrum.} \quad (3.9)$$

For ψ in the cuspidal basis, let it_ψ denote the archimedean parameter $\mu_{\pi,1}$ of the representation π containing ψ; the second inner product $\langle\psi, P_h(z,s)\rangle$ equals the Fourier coefficient of ψ, $\overline{\rho_\psi(-h)}$ times a factor bounded by $(1+|t_\psi|)^B e^{\frac{\pi}{2}|t_\psi|}$. The Fourier coefficient $\overline{\rho_\psi(-h)}$ is bounded by $O(|h|^{\theta+o(1)})$ by Hypothesis H_θ at the non-archimedean places. The problem now, as was pointed out by Selberg, is to have a bound for the triple product integral $\langle(\ell_1 y)^{k/2}\varphi_2(\ell_1 z).(\ell_2 y)^{k/2}\overline{\varphi_2}(\ell_2 z),\ \psi\rangle$ which exhibits an exponential decay for the form $O((1+|t_\psi|)^C e^{-\frac{\pi}{2}|t_\psi|})$, so as to compensate the exponential growth of $\langle\psi, P_h(z,s)\rangle$. In this generality, this exponential decay property for triple product was achieved by Sarnak in [69]; later, a representation theoretic version of Sarnak's arguments as well as some improvements were given by Bernstein–Reznikov [2]. The final consequence of these bounds is the following estimate

$$\Sigma_-(\varphi_2, \ell_1, \ell_2, h) = O_{\varphi_2,\varepsilon}\big((\ell_1\ell_2)^A q^{1/2+\theta+\varepsilon}\big). \quad (3.10)$$

This approach is important for several reasons:

- It ties more closely the subconvexity problem for GL_2 L-functions – a problem whose origin lies in analytic number theory – to the Ramanujan–Petersson conjecture for $\mathrm{GL}_2(\boldsymbol{A}_{\mathbb{Q}})$; or, in other words, to the spectral gap property which is a classical problem in the harmonic analysis of groups;

- it gives an hint that automophic period integrals might be useful in the study of the subconvexity problem: this will be largely confirmed in Section 4.

- This approach is sufficiently smooth that it can be extended to number fields of higher degree: a few years ago, Cogdell–Piatetski-Shapiro–Sarnak used the amplification method in conjunction with this approach to obtain (3.2) when F is totally real and π_∞ is a holomorphic discrete series (see [13]).

Remark 3.2. The methods of sections 3.2 and 3.3 are closely related. This can be seen already by remarking that Weil's bound for Kloosterman sums yield the saving $q^{3/4+\varepsilon}$ in (3.8) which is precisely the saving following from Hypothesis $H_{1/4}$ in (3.10); moreover $H_{1/4}$ (a.k.a the Selberg–Gelbart–Jacquet bound) can be obtained by applying Weil's bound to the Kloosterman sums.One can push this coincidence further, by applying, in (3.7) the Kuznetzov–Petersson formula *backwards* in order to transform the sums of Kloosterman sums into sums of Fourier coefficients of Maass

forms:

$$(3.7) = \sum_{\psi}\sum_{h'}\Big(\sum_{\mp\ell_1 n-\ell_2 m=h'}\alpha_m\overline{\rho_{\varphi_2}(m)}\beta_n\rho_{\varphi_2}(n)\Big)\overline{\rho}_\psi(-h)\rho_\psi(h')\tilde{\mathcal{V}}(h,h',it_\psi)$$
$$+\text{ Discrete series Spectrum }+\text{ Eisenstein spectrum.}\quad(3.11)$$

Thus, we have realized the spectral expansion of the shifted convolution sum $\Sigma_\pm(\varphi_2,\ell_1,\ell_2,h)$ in a way similar to that obtained in (3.9); from there, we may use again the full force of spectral theory. This may look like a rather circuituous path to obtain the spectral expansion; this method however has some technical advantage over the method discussed in Section 3.3: it works even if φ_2 is a Maass form, without the need to find appropriate test vector or to obtain exponential decay for triple product integrals ! The spectral decomposition (3.11) will be very useful in the next section.

3.4. The case of a varying central character. The methods discussed so far are sufficient to establish (3.2), (3.3), (3.4) as long as the conductor of the central character, ω_1 say, of π_1 is significantly smaller than q_1. The case of a varying central character reveals new interesting features which we discuss here. To simplify, we consider the extremal (in a sense hardest) case where both conductors are equal $q_{\omega_1}=q_1=:q$.

3.4.1. Subconvexity via bilinear Kloosterman fractions. As usual for the subconvexity problem, the first result is due to Duke–Friedlander–Iwaniec for the case (3.3) [22], [23]. As pointed out above, the problem of bounding $L(\pi_1,s)$ for $\Re es=1/2$ may be formulated as the problem of bounding a Rankin–Selberg L-function

$$L(\pi_1\times\pi_2,s)=L(\pi_1,s)^2$$

where $\pi_2=1\boxplus 1$ is the representation corresponding to the the fully unramified Eisenstein series. Eventually, another approach was considered in [22], [23], which comes from the identity

$$|L(\pi_1,s)|^2=L(\pi_1\times\overline{\chi},1/2)L(\tilde{\pi}_1\times\chi,1/2)$$

where $\tilde{\pi}_1$ is the contragredient and $\chi=\omega_1|.|^{-it}$, $t=\Im ms$. The amplification method applied to the family $\{L(\pi\times\overline{\chi},1/2)L(\tilde{\pi}\times\chi,1/2),\ q_\pi=q_1:=q,\omega_\pi=\omega_1\}$ yields in practice to shifted convolution sums of the form ([22], [23])

$$\sum_{\ell_1 ad-\ell_2 bc=h}\overline{\chi}(a)\chi(c)\mathcal{W}\Big(\frac{a}{q^{1/2}},\frac{b}{q^{1/2}},\frac{c}{q^{1/2}},\frac{d}{q^{1/2}}\Big),$$

with $h\approx q$, $h\equiv 0(q)$. The later is essentially a truncated version of the shifted convolution sums associated to the Eisenstein series $E(1,\chi)$ of the representation $1\boxplus\chi$; the new feature by comparison with the previous shifted convolution problems

is that the coefficients $\rho_{E(1,\chi)}(n) = \sum_{bc=n} \chi(c)$ vary with q, which is essentially the range of the variables $m = ad$ and $n = bc$. Since χ has conductor q and a, c vary in ranges of size $\approx q^{1/2}$ one cannot really use the arithmetical structure of the weights $\overline{\chi}(a),\ \chi(c)$ so this shifted convolution problem is basically reduced to the non-trivial evaluation of a quite general sum:

$$\sum_{\ell_1 ad - \ell_2 bc = h} \alpha_a \gamma_c \mathcal{W}\Big(\frac{a}{q^{1/2}}, \frac{b}{q^{1/2}}, \frac{c}{q^{1/2}}, \frac{d}{q^{1/2}}\Big) = MT((\alpha_a), (\gamma_c), \ell_1, \ell_2, h) + O((\ell_1\ell_2)^A q^{1-\delta}) \tag{3.12}$$

for some $\delta > 0$ absolute and where $MT((\alpha_a), (\gamma_c), \ell_1, \ell_2, h)$ denotes a natural main term and with $(\alpha_a)_{a\sim q^{1/2}}$, $(\gamma_c)_{c\sim q^{1/2}}$ arbitrary complex numbers of modulus bounded by 1. Since the b variable is smooth, the condition $\ell_1 ad - \ell_2 bc = h$ is essentially equivalent to the congruence condition $\ell_1 ad \equiv h \bmod c\ell_2$. One can then analyze this congruence by Poisson summation applied on the remaining smooth variable d which yields sums of Kloosterman fractions of the shape

$$\sum_{\substack{a\sim A, c\sim C\\ (a,c)=1}} \alpha_a \gamma_c e\Big(h\frac{\bar{a}}{c}\Big), \quad \text{for } h \neq 0$$

and where the values of a, c, h and α_a, γ_c may be different from the previous ones. In [21] such sums are bounded non-trivially for any ranges A, C (the most crucial one being $A = C$).

A remarkable feature of this proof is that the bound is obtained from an application of the amplification method in a very unexpected direction, namely by amplifying the trivial (!) multiplicative characters $\chi_{0,a}$ of modulus a in the family of sums

$$\Big\{ \sum_{\substack{c\sim C\\ (a,c)=1}} \gamma_c \chi(c) e\Big(h\frac{\bar{a}}{c}\Big),\ \chi \bmod a,\ a \sim A\Big\}.$$

Remark 3.3. Note that (3.12) is more general than needed for (3.3) and may be used in other contexts (e.g. Bombieri–Vinogradov type results). On the other hand, in the subconvexity context, this method uses the special shape of Eisenstein series and does not seem to generalize to Rankin–Selberg L-functions.

3.4.2. Subconvexity of Rankin–Selberg L-functions via subconvexity for twisted L-functions. The case of Rankin–Selberg L-functions over $\mathbb{Q}$, $L(\pi_1 \times \pi_2, s)$ when π_2 is essentially fixed and π_1 has a central character ω_1 of large conductor was treated in [39], [60]. In the case of a varying central characters, subconvexity comes from an estimate for an average of shifted convolution sums of h of the form:

$$\sum_{0<|h|\ll q} \overline{\omega}(h) \Sigma_{\pm}(\varphi_2, \ell_1, \ell_2, h) \ll_{\varphi_2} (\ell_1\ell_2)^A q^{3/2-\delta} \tag{3.13}$$

for some $A, \delta > 0$ absolute. Observe however that this is stronger than just the shifted convolution problem on average over h. In particular even under the Ramanujan–Petersson conjecture (H_0), the individual bound (3.10) is "just" not sufficient: this means that one has to account for the averaging over the h variable.

This bound is achieved through the spectral decomposition of the shifted convolution sums (3.11): plugging this formula into the left-hand side of (3.13) one obtains a sum over the orthonormal basis $\{\psi\}$ of sums of the form

$$\sum_{0<|h|\ll q} \overline{\omega}(h)\rho_\psi(-h)$$

if ψ belong to the space V_τ of some automorphic representation, the later sums are partial sums associated to the twisted L-function $L(\tau \times \overline{\omega}, s)$. In that case, the subconvexity bound for twisted L-functions (3.2) is exactly sufficient to give (3.13).

Remark 3.4. Hence the subconvexity bound for an L-functions of degree 4 has been reduced to a collection of subconvex bounds for L-functions of automorphic forms of small level twisted by the original central character ω ! This surprising phenomenon is better explained via the approach described in the next section.

4. Subconvexity via periods of automorphic forms

4.1. The various perspectives on an L-function. From the perspective of analytic number theory, the definition of L-function might be "an analytic function sharing the key features of $\zeta(s)$: analytic continuation, functional equation, Euler product."

However, there are various "incarnations" of L-functions attached to automorphic forms; although equivalent, different features become apparent in different incarnations. For instance, one can define and study L-functions via constant terms of Eisenstein series (the Langlands-Shahidi method), via periods of automorphic forms (the theory of integral representations, which begins with the work of Hecke, or indeed already with Riemann), or via a Dirichlet series (which is often taken as their defining property).

Thus far in this article, we have discussed the subconvexity from the perspective of Dirichlet series. In particular, we have studied periods (e.g. (2.4)) by relating them to L-functions (via (2.2)) and then proving subconvexity for the latter. Relatively recently, the subconvexity question has also been succesfully approached via the "period" perspective by reversing this usual process: namely by deducing subconvexity from a geometric study of the periods. The first such result (in the eigenvalue aspect) was given by Bernstein-Reznikov [3], [4], and a little later a result in the level aspect was given by Venkatesh [75]. The two methods seem to be quite distinct. We shall discuss these briefly, and then discuss in more detail the joint work [61] of the authors, which also uses the period perspective.

These approaches are closely related to existing work, but in many cases the period perspective allows certain conceptual simplifications and it brings together harmonic analysis and ideas from dynamics. Such conceptual simplifications are particularly of value in passing from $\mathbb{Q}$ to a general number field; so far, with the exception of the result of Cogdell–Piatetski-Shapiro–Sarnak, all the results in Theorem 6 in the case $F \neq \mathbb{Q}$ are proven via the period approach.

On the other hand, it might be noted that a slight drawback to the period approach to subconvexity is that, especially for automorphic representations with complicated ramification, one must face the difficulty of choosing appropriate test vectors.

4.2. Triple product period and triple product L-function. At present, all known results towards the subconvexity of triple product L-functions $L(\pi_1 \times \pi_2 \times \pi_3, 1/2)$ arise from the "period" perspective.

The period of interest is

$$\int_{\mathrm{PGL}_2(\mathbb{Q})\backslash \mathrm{PGL}_2(\boldsymbol{A})} \varphi_1(g)\varphi_2(g)\varphi_3(g)\, dg$$

where $\varphi_i \in \pi_i$, and each π_i is an automorphic cuspidal representation of GL_2. It is expected that this period, and the variants when GL_2 is replaced by the multiplicative group of a quaternion algebra, is related to the central value of the triple product L-function $L(\pi_1 \times \pi_2 \times \pi_3, 1/2)$, see [40]. A precise relationship has been computed for the case of Maass forms at full level in [77]; indeed, the following formula is established:

$$\left| \int_{\mathrm{SL}_2(\mathbb{Z})\backslash \mathbb{H}} \varphi_1\varphi_2\varphi_3 d\mu \right|^2 = \frac{\Lambda(\varphi_1 \times \varphi_2 \times \varphi_3, 1/2)}{\Lambda(\wedge^2\varphi_1, 1)\Lambda(\wedge^2\varphi_2, 1)\Lambda(\wedge^2\varphi_3, 1)} \tag{4.1}$$

where Λ denotes the completed L-function and $d\mu$ is a suitable multiple of $\frac{dx\,dy}{y^2}$.

4.2.1. The eigenvalue aspect: the method of Bernstein–Reznikov. Let Γ be a (discrete) cocompact subgroup of $\mathrm{SL}_2(\mathbb{R})$, let $\mathbb{H}$ be the upper half-plane, let φ_1, φ_2 be fixed eigenfunctions of the Laplacian on $\Gamma\backslash\mathbb{H}$ and φ_λ an eigenfunction with eigenvalue $\lambda := 1/4+r^2$. In the paper [4], Bernstein and Reznikov establish the following bound:

$$r^2 e^{\pi r/2} \left| \int_{\Gamma\backslash\mathbb{H}} \varphi_1(z)\varphi_2(z)\varphi_\lambda(z)\, d\mu_z \right|^2 \ll_\varepsilon r^{5/3+\varepsilon}. \tag{4.2}$$

In fact the bound (4.2) remains valid if $\Gamma\backslash\mathbb{H}$ has only finite volume and $\varphi_1, \varphi_2, \varphi_\lambda$ are cusp forms. In particular, when $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ and $\varphi_1, \varphi_2, \varphi_\lambda$ are Hecke–Maass forms associated to automorphic representations $\pi_1, \pi_2, \pi_\lambda$ respectively, the bound (4.2) translates, via (4.1), to the subconvex bound

$$L(\pi_1 \times \pi_2 \times \pi_\lambda, 1/2) \ll_{\varepsilon,\pi_1,\pi_2} r^{5/3+\varepsilon} \tag{4.3}$$

while the convexity bound for the left-hand side is $r^{2+\varepsilon}$.

Their method is based on the properties of the (local) real group $G = \mathrm{PGL}_2(\mathbb{R})$ and, in particular, on the fact that the space of G-invariant functionals on $\pi_1 \otimes \pi_2 \otimes \pi_\lambda$ is *at most* one dimensional. Hence the proof is purely local and by contrast to the method of [75], does not use either Hecke operators or the spectral gap.

4.2.2. The level aspect: the method of Venkatesh. Let F be a number field. Let π_2, π_3 be fixed automorphic cuspidal representations on $\mathrm{PGL}_2(\boldsymbol{A}_F)$ – say with coprime conductor – and let π_1 be a third automorphic cuspidal representation with conductor $\mathfrak{q}$, a prime ideal of F. In [75] it is established that

$$L(\pi_1 \times \pi_2 \times \pi_3, 1/2) \ll_{\pi_{1,\infty},\pi_2,\pi_3} \mathrm{N}(\mathfrak{q})^{1-\frac{1}{13}} \tag{4.4}$$

contingent on a suitable version of (4.1) when the level of one factor varies.[4] The convexity bound for the left-hand side is $\mathrm{N}(\mathfrak{q})^{1+\varepsilon}$.

Remark 4.1. In [75], a form of (4.4) is proved when π_2 and/or π_3 are Eisenstein series: in that case, (4.1) corresponds to simple computations in the Rankin–Selberg method and so is unconditional. In particular, this yields the bounds (3.3) and (3.4) for π_1 with trivial central character.

For reasons of space, we do not explain the details of the proof; in any case, this can also be approached by the method outlined in Section 4.3. It uses, in particular, quantitative results and ideas from ergodic theory, and the bound H_θ with $\theta < 1/4$, in the notation of Section 3.3.

4.3. Central character. In this section, we return to Section 3.4 and explain, via periods, the bound (3.4). In particular, this sheds light on the "reason" for the reduction to a lower rank subconvexity problem that was encountered in that section. The content of this section is carried out in detail in [61].

Let π_1, π_2 be automorphic cuspidal representations of $\mathrm{GL}_2(\boldsymbol{A}_F)$. Let ω be the central character of π_1. For simplicity, we restrict ourselves to the case where π_2, the "fixed" form, has level 1 and trivial central character; and where "all the ramification of π_1 comes from the central character," i.e. π_1 and ω have the same conductor $\mathfrak{q}$.

Let us first give a very approximate "philosophical" overview of the proof. There is an identity between mean values of L-functions of the following type:

$$\sum_{\pi_1} L(\pi_1 \times \pi_2, 1/2) \longleftrightarrow \sum_{\tau \text{ level } 1} L(\tau, 1/2) L(\tau \times \omega, 1/2) \tag{4.5}$$

where the left-hand summation is over π_1 of central character ω and conductor $\mathfrak{q}$, whereas the right-hand summation is over automorphic representations τ of trivial

[4] This has not appeared in the literature to our knowledge, except in the case where one of the π_j are Eisenstein; however, it should amount to a routine though very involved computation of p-adic integrals.

central character and level 1. It includes the trivial (one-dimensional) automorphic representation, which is in fact the dominant term and actually needs to be handled by regularization. [5]

By means of a suitable amplifier, one can restrict the left-hand summation to pick out a given π_1. When one does this, the necessary bounds on the right-hand side follow from two different inputs:

1. Subconvexity for $L(\tau \times \omega, 1/2)$ (in the aspect where ω varies), to handle the nontrivial τ.

2. A bound showing decay of matrix coefficients of p-adic groups, to handle the contribution of τ the trivial representation

4.3.1. The source of (4.5) via periods. Writing $Y_{\boldsymbol{A}} = \mathrm{PGL}_2(F)\backslash\mathrm{PGL}_2(\boldsymbol{A}_F)$, we note that the Rankin–Selberg L-function may be expressed as a period integral:

$$L(s, \pi_1 \times \pi_2) \sim \int_{Y_{\boldsymbol{A}}} \varphi_1(g)\varphi_2(g)E_s(g)\, dg$$

where $\varphi_i \in \pi_i$ are the respective newforms, and E_s is the Eisenstein series corresponding to the new vector of the automorphic representation $|\cdot|^s \boxplus \omega^{-1}|\cdot|^{-s}$. Here $\sim$ means that there is a suitable constant of proportionality, depending on the archimedean types of the representations.

Let $\mathcal{B}_{\omega,\mathfrak{q}}$ be an orthogonal basis for the space of forms on $\mathrm{GL}_2(F)\backslash\mathrm{GL}_2(\boldsymbol{A}_F)$ of level $\mathfrak{q}$ and central character ω; let $\mathcal{B}_{1,1}$ be an orthogonal basis for the space of forms on $Y_{\boldsymbol{A}}$ of full level and trivial central character. By spectral expansion, we have the following identity:

$$\begin{aligned}\sum_{\varphi_1 \in \mathcal{B}_{\omega,\mathfrak{q}}} \left|\int_{Y_{\boldsymbol{A}}} \varphi_1\varphi_2 E_s\right|^2 &= \int_{Y_{\boldsymbol{A}}} |\varphi_2 . E_s|^2 \\ &= \int_{Y_{\boldsymbol{A}}} |\varphi_2|^2\, |E_s|^2 = \sum_{\psi \in \mathcal{B}_{1,1}} \langle |E_s|^2, \psi\rangle \overline{\langle |\varphi_2|^2, \psi\rangle}.\end{aligned} \tag{4.6}$$

Here the ψ- summation is, *a priori*, over an orthonormal basis for $L^2(Y_{\boldsymbol{A}})$; however, the summand $\langle |\varphi_2|^2, \psi\rangle$ vanishes unless ψ is of level 1 and trivial central character. Note that the ψ-summation should, strictly, include a continuous contribution for the Eisenstein series, which also needs to be suitably regularized. This is not a trivial matter and occupies a good deal of [61]; we shall supress it for now.

In any case, if ψ belongs to the space of an automorphic representation τ, then the Rankin–Selberg method shows that $\langle |E_s|^2, \psi\rangle$ is a multiple of $L(\tau, 2s - 1/2)$ $L(\tau \times \omega, 1/2)$. Thus (4.6) basically yields (4.5)!

[5] Note that "morally", when τ is trivial, the L-function $L(\tau, s)L(\tau \times \omega, s) = \zeta(s+1/2)\zeta(s-1/2)L(\omega, s-1/2)$ $L(\omega, s + 1/2)$. Thus we obtain a pole at $s = 1/2$.

4.3.2. Amplification and the decay of matrix coefficients. We restrict to $s = 1/2$ for concreteness, although the method works for any s. The identity (4.6) does not suffice to obtain a nontrivial bound on $\int_{Y_A} \varphi_1\varphi_2 E_{1/2}$, for the left-hand summation is too large. To localize it, one introduces an amplifier. We phrase it adelically, but it should be made clear this is still the amplifier of Friedlander–Iwaniec.

For any function f on $\mathrm{GL}_2(F)\backslash\mathrm{GL}_2(\boldsymbol{A}_F)$ and any $g_0 \in \mathrm{GL}_2(\boldsymbol{A}_F)$, we write $f^{g_0}(g) = f(gg_0^{-1})$. Then one has the following tiny variant of (4.6), for $g_1, g_2 \in \mathrm{GL}_2(\boldsymbol{A}_F)$:

$$\begin{aligned}\sum_{\varphi_1\in\mathcal{B}_{\omega,\mathfrak{q}}} \left(\int_{Y_A} \varphi_1^{g_1^{-1}}\varphi_2 E_{1/2}\right)\overline{\left(\int_{Y_A} \varphi_1^{g_2^{-1}}\varphi_2 E_{1/2}\right)} \\ = \sum_{\psi\in L^2(Y_A)} \langle E_{1/2}^{g_1}\overline{E_{1/2}^{g_2}}, \psi\rangle\overline{\langle\overline{\varphi_2^{g_1}}\varphi_2^{g_2}, \psi\rangle}.\end{aligned} \tag{4.7}$$

This is again an identity of this shape (4.5), but with slightly more freedom due to the insertion of g_1, g_2. The left-hand (resp. right-hand) side is still proportional, by the Rankin–Selberg method, to $L(\pi_1 \times \pi_2, 1/2)$ (resp. $L(\tau, 1/2)L(\tau \times \omega, 1/2)$, if $\psi \in \tau$) but the constants of proportionality depend – in a precisely controllable way – on g_1, g_2. In effect, this allows one to introduce a "test function" $h(\pi_1)$ into the identity (4.5), thereby shortening the effective range of summation. It should be noted that in (4.7), by contrast with (4.5), the right hand ψ-summation is no longer over ψ of level 1; however, it involves only those ψ which are invariant by $\mathrm{PGL}_2(\hat{\mathbb{Z}}) \cap g_1^{-1}\mathrm{PGL}_2(\hat{\mathbb{Z}})g_1 \cap g_2^{-1}\mathrm{PGL}_2(\hat{\mathbb{Z}})g_2$, and in particular their level is bounded in a way that depends predictably on g_1, g_2.

A subconvex bound for $L(\pi_1 \times \pi_2, 1/2)$ follows from any method to get nontrivial bounds on the right-hand side of (4.7) for general g_1, g_2.

1. To deal with the case when ψ is perpendicular to the constants, we note that the terms $\langle E_{1/2}^{g_1}E_{1/2}^{g_2}, \psi\rangle$ are, by Rankin–Selberg, certain multiples of $L(\tau, 1/2)L(\tau \times \omega, 1/2)$ whenever $\psi \in \tau$, the space of an automorphic representation. We then apply subconvex bounds for $L(\tau \times \omega, 1/2)$, in the aspect when ω varies.
2. To deal with the case $\psi = \mathrm{Const}$, we note that the term $\langle \varphi_2^{g_1}\overline{\varphi_2^{g_2}}, \psi\rangle$ is, in that case, simply a multiple of the matrix coefficient $\langle \varphi_2^{g_1g_2^{-1}}, \varphi_2\rangle$. Thus it is bounded by bounds on the decay of matrix coefficients.

Obviously this description is dishonest, for the term $\langle |E_{1/2}|^2, \psi\rangle$ is not even convergent for $\psi = \mathrm{Const}$! However, this is a technical and not a conceptual difficulty: the only two analytic ingredients required are the two above.

4.3.3. Mysterious identities between families of L-functions. In a sense, the period identity (4.6) (or, approximately equivalent, the identity (4.5)) is the key point of

the above discussion; it explains immediately why one has the "reduction of degree" discussed in Section 3.4. This is another example of the phenomenon discussed in Section 4.1: the identity (4.5) is obvious from the "period" perspective, but not at all clear from the viewpoint of L-functions considered as Dirichlet series.

Another example of such a phenomenon – identities between *a priori* different families of L-functions – is Motohashi's beautiful formula [64] for the 4th moment of ζ. Roughly speaking, it relates integrals of $|\zeta(1/2+it)|^4$ to sums of $L(\varphi, 1/2)^3$, where φ varies over Maass forms. If φ is a suitably normalized Maass form on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, its completed L-function is given by the Hecke period $\Lambda(\varphi, 1/2+it) = \int_0^\infty \varphi(iy) y^{it} d^\times y$. Applying Plancherel's formula shows that $\frac{1}{2\pi}\int_{-\infty}^{\infty} |\Lambda(\varphi, 1/2+it)|^2 dt = \int_{y=0}^{\infty} |\varphi(iy)|^2 d^\times y$. Again, one can spectrally expand $|\varphi|^2$, yielding:

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty} |\Lambda(\varphi, 1/2+it)|^2\, dt &= \sum_{\psi} \frac{\langle |\varphi|^2, \psi\rangle}{\langle \psi, \psi\rangle} \int_0^\infty \psi(iy)\, d^\times y \\ &= \sum_{\psi} \frac{\langle |\varphi|^2, \psi\rangle}{\langle \psi, \psi\rangle} \Lambda(\psi, 1/2)\end{aligned} \tag{4.8}$$

where, again, the ψ-sum is over an orthogonal basis, suitably normalized, for $L^2(Y_0(1))$, and, again, we suppress the continuous spectrum; hence (4.8) expresses a relation between mean values of $L(\varphi, 1/2+it)$, where t varies, and the family of L-functions $L(\psi, 1/2)$, where ψ varies. Specializing (4.8) to the case of φ the Eisenstein series at the center of symmetry yields a formula "of Motohashi type." We emphasize that this argument has not been carried out rigorously to our knowledge and it would likely involve considerable technical difficulty (for the integrals diverge in the eisenstein case). Nevertheless, this approach may have value insofar as it offers some insight into the origin of such formulae. A. Reznikov has given a very general and elegant formalism [67] that encapsulates such identities as (4.5) and (4.8); one hopes that further analytic applications will stem from his formalism.

5. Applications

5.1. Subconvexity and functoriality. Via the functoriality principle of Langlands, it is now understood that the same L-function may be attached to automorphic forms on different groups. This gives rise to the possibility of studying the same L-function in different ways.

A recent instance where this kind of idea played a decisive role was the attempt to solve the subconvexity problem for the L-functions of the class group characters of a quadratic field K of large discriminant. In [24], the problem was solved but only under the assumption that K has sufficiently many small *split* primes (this would follow from GRH, but so far, has been established unconditionally only for special discriminants). This assumption, which was also encountered by the second author

in [75] in the context of periods and is closely related to Linnik's condition, is a fundamental and major unsolved issue that arises in many contexts, e.g. in work on the André–Oort conjecture [79]. A key observation of [23], is that, by functoriality, (in that case due to Hecke and Maass) a class group character L-function is the L-function of a Maass form of weight 0 or 1, with Laplace eigenvalue $1/4$. For these, as we have just seen, the subconvexity problem can be solved independently of any assumption.

In view of this example, we find it useful to spell out explicitly some direct consequences of the subconvex bounds of Theorem 6 and of functoriality.

Corollary 5.1. *Let F be a fixed number field and $\rho\colon \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\mathbb{C})$ be a modular Galois representation (for instance, if the image of ρ in $\mathrm{PGL}_2(\mathbb{C})$ is soluble). Let $\mathfrak{q}_\rho$ be the Artin conductor of ρ and let $L(\rho, s)$ be its Artin L-function, then for $\Re es = 1/2$*

$$L(\rho, s) \ll_{F,s} N_{F/\mathbb{Q}}(\mathfrak{q}_\rho)^{1/4-\delta}$$

for $\delta > 0$ some absolute constant.

Corollary 5.2. *Let F be a fixed number field and K be an extension of F of absolute discriminant $\mathrm{disc}(K/\mathbb{Q}) =: \Delta_K$ and let $\zeta_K(s)$ be the Dedekind zeta function of K; then, if K/F is abelian or cubic, one has for $\Re es = 1/2$*

$$\zeta_K(s) \ll_{F,s} |\Delta_K|^{1/4-\delta}$$

for $\delta > 0$ some absolute constant.

Corollary 5.3. *Let F be a fixed number field, π be a fixed $\mathrm{GL}_2(\boldsymbol{A}_F)$-automorphic cuspidal representation and let K be an extension of F of absolute discriminant $\mathrm{disc}(K/\mathbb{Q}) =: \Delta_K$. If K/F is abelian or cubic, we denote by π_K the base change lift of π from F to K (which exist by the works of Saito–Shintani–Langlands and Jacquet–Piatetski-Shapiro–Shalika). For $\Re es = 1/2$, one has*

$$L(\pi_K, s) \ll_{F,\pi,s} |\Delta_K|^{1/2-\delta}$$

for $\delta > 0$ some absolute constant.

5.2. Equidistribution on quaternionic varieties. We define a quaternionic variety as the locally homogeneous space given as an adelic quotient of the following form: for F a totally real number field, B a quaternion algebra over F, let $\boldsymbol{G}$ be the $\mathbb{Q}$-algebraic group $\mathrm{res}_{F/\mathbb{Q}}\mathrm{B}^\times/Z(\mathrm{B}^\times)$; one has

$$G(\mathbb{R}) \simeq \mathrm{PGL}_2(\mathbb{R})^{f'} \times \mathrm{SO}(3, \mathbb{R})^{f-f'}$$

where $f = \deg F$ and f' is the number of real place of F for which B splits. Let K_∞ be a compact subgroup of $\boldsymbol{G}(\mathbb{R})$ of the form

$$\mathrm{SO}(2, \mathbb{R})^{f'} \times \prod_{v=1}^{f-f'} \mathrm{K}_v$$

with $\mathrm{K}_v =$ either $\mathrm{SO}_2(\mathbb{R})$ or $\mathrm{SO}_3(\mathbb{R})$ and let X denote the quotient $\boldsymbol{G}(\mathbb{R})/\mathrm{K}_\infty$; finally let K_f be an open compact subgroup of $\boldsymbol{G}(\boldsymbol{A}_f)$ and $\mathrm{K} := \mathrm{K}_\infty.\mathrm{K}_f$.

The quaternionic variety $V_{\mathrm{K}}(\boldsymbol{G}, X)$ is defined as the quotient

$$V_{\mathrm{K}}(\boldsymbol{G}, X) := \boldsymbol{G}(\mathbb{Q})\backslash X \times \boldsymbol{G}(\boldsymbol{A}_f)/\mathrm{K}_f = \boldsymbol{G}(\mathbb{Q})\backslash \boldsymbol{G}(\boldsymbol{A}_{\mathbb{Q}})/\mathrm{K}.$$

It has the structure of a Riemannian manifold whose connected components are quotients, by a discrete subgroup of $\boldsymbol{G}(\mathbb{R})$, of the product of $(\mathbb{H}^{\pm})^{f'} \times (S^2)^{f''}$ for $f'' \leqslant f - f'$. The case of the sphere and of the modular surface correspond to the case $F = \mathbb{Q}$, B the algebra of 2×2 matrices $M_2(\mathbb{R})$ or the Hamilton quaternions $\mathrm{B}^{(2,\infty)}$.

Let K/F be a quadratic extension with an embedding into B, and let $\boldsymbol{T}$ denote the $\mathbb{Q}$-torus "$\mathrm{res}_{F/\mathbb{Q}} K^\times/F^\times$". As was pointed out in Section 2.2.1, there exists, in great generality, a precise relationship between

1. central values of some Rankin–Selberg L-function $L(\pi_\chi \times \pi_2, s)$ (for which the sign of the functional equation $w(\pi_\chi \times \pi_2)$ is $+1$); and
2. (the square of) twisted Weyl sums

$$\int_{\boldsymbol{T}(\mathbb{Q})\backslash \boldsymbol{T}(\boldsymbol{A}_{\mathbb{Q}})} \chi(t)\varphi_2(z.t)\, dt.$$

These Weyl sums describe the distribution properties of toric orbits, $\boldsymbol{T}(\mathbb{Q})\backslash z.\boldsymbol{T}(\boldsymbol{A}_{\mathbb{Q}})$ of cycles associated to (orders of) K inside $V_{\mathrm{K}}(\boldsymbol{G}, X)$.

The general scheme is that, in cases where these formula have been written out explicitly, the subconvex bound (3.2) (along possibly with hypothesis H_θ for some $\theta < 1/2$) yields at once the equidistribution of the *full* orbit and the subconvex bounds (3.4) yield the equidistribution of *big enough* suborbits of the toric orbit. We present below some sample results on these lines:

5.2.1. Hilbert's eleventh problem. When B is totally definite, $\mathrm{K}_\infty = \mathrm{SO}_2(\mathbb{R})^f$, $X = (S^2)^f$ is a product of spheres. In this case, the equidistribution of toric orbits (relative to a totally imaginary quadratic field) above can be interpreted in terms of the integral representations of a totally positive integer $d \in \mathcal{O}_F$ by a totally positive definite quadratic form q (more precisely $-q$ "is" the norm form $N_{B/F}(\mathbf{x})$ on the space of quaternions of trace 0). The following theorem of Cogdell–Piatetski-Shapiro–Sarnak combines the formula of [1] with (3.2) for π_2 holomorphic.

Theorem 7. *Let F be a totally real number field and q be an integral positive definite quadratic form over F; there is an absolute (ineffective) constant $N_{F,q} > 0$ such that if d is a squarefree totally positive integer with $N_{F/\mathbb{Q}}(d) > N_F$ then d is integraly represented by q iff d is everywhere locally integrally represented. Moreover, in the later case, the number, $r_q(d)$, of all such integral representation satisfies*

$$r_q(d) \gg_{q,F} N_{F/\mathbb{Q}}(d)^{1/2+o(1)} \quad \text{as } N_{F/\mathbb{Q}}(d) \to +\infty.$$

Remark 5.1. The question of the integral representability of d by some form in the genus of q was completely settled a long time ago by Siegel, in a quantitative way, through the Siegel mass formula. The present theorem (in a slightly more precise form) can then be interpreted by saying that the various representations d are *equidistributed* amongst the various genus classes of q; moreover it can be strengthened to an "equidistribution on ellipsoids" statement, cf. [26] for $F = \mathbb{Q}$.

5.2.2. CM points on quaternionic Shimura varieties. When B is indefinite at some real place and $\mathrm{K}_\infty = \mathrm{SO}_2(\mathbb{R})^{f'} \times \mathrm{SO}_3(\mathbb{R})^{f-f'}$ the quaternionic variety $V_{\mathrm{K}}(\boldsymbol{G}, X)$ is a Shimura variety, $\mathrm{Sh}_{\mathrm{K}_f}(\boldsymbol{G}, X)$ (a Hilbert modular variety of complex dimension f'). It has the structure of the complex points of an algebraic variety defined over some reflex field E/F.

In this setting, the generalization of the set of Heegner point is the so called set of "CM" points, $\mathcal{H}_{\mathfrak{d}}$, which is associated to a quadratic order $\mathcal{O}_{\mathfrak{d}}$ (say of discriminant $\mathfrak{d}$) of a (not necessarily fixed) totally imaginary K/F. In that case and under some natural local condition, the equidistribution of

$$\mathcal{H}_{\mathfrak{d}} = \boldsymbol{T}(\mathbb{Q})\backslash z_{\mathfrak{d}}.\boldsymbol{T}(\boldsymbol{A}_{\mathbb{Q}})/\boldsymbol{T}(\widehat{\mathcal{O}}_{\mathfrak{d}})$$

on $\mathrm{Sh}_{\mathrm{K}_f}(\boldsymbol{G}, X)$ as $|N_{F/\mathbb{Q}}(\mathfrak{d})| \to +\infty$ was established independently by Clozel–Ullmo, Cohen and Zhang [12], [14], [82] by using the subconvex bound (3.2) of the second author. For instance, one has

Theorem 8. *Suppose* $\mathrm{K}_f = \mathrm{K}_{f,\max}$ *is a maximal compact subgroup of* $\boldsymbol{G}(\boldsymbol{A}_F)$*, then for* $|N_{F/\mathbb{Q}}(\mathfrak{d})| \to +\infty$ *and* $\mathfrak{d}$ *coprime with* $\mathrm{disc}(F)$*, the set* $\mathcal{H}_{\mathfrak{d}}$ *becomes equidistributed on* $\mathrm{Sh}_{\mathrm{K}_f}(\boldsymbol{G}, X)$ *w.r.t. the hyperbolic measure.*

Similarly, as in Theorem 5, the bound (3.4) allows one to show the equidistribution of strict suborbits of $z_{\mathfrak{d}}$:

Theorem 9. *With the notations as above, there is an absolute constant* $0 < \eta < 1$ *such that, for any subtoric orbit* $\mathcal{H}'_{\mathfrak{d}} \subset \mathcal{H}_{\mathfrak{d}}$ *of size satisfying* $|\mathcal{H}'_{\mathfrak{d}}| \geqslant |\mathcal{H}_{\mathfrak{d}}|^{\eta}$*, then* $\mathcal{H}'_{\mathfrak{d}}$ *is equidistributed on* $\mathrm{Sh}_{\mathrm{K}_f}(\boldsymbol{G}, X)$ *as* $|N(\mathfrak{d})| \to +\infty$.

As was pointed out by Zhang [82], the possibility of considering strict suborbits of the full toric orbit has a nice arithmetic interpretation; the Galois orbits on CM points correspond to "subtoric orbits" of the type considered in Theorem 9.

6. Linnik's ergodic method: a modern perspective

As discussed, Linnik achieved partial results towards Theorems 1–3 by using some ingenious ideas which he collectively referred to as "the ergodic method." As Linnik pointed out (see, e.g., [57, Chapter XI, comments on Chapters IV–VI]) despite this name, this method remained rather *ad hoc* and did not fit into ergodic theory as it is

normally understood: that is to say, dynamics of a measure-preserving transformation. The joint work of the authors with M. Einsiedler and E. Lindenstrauss [29], remedies this, both putting Linnik's original work into a more standard ergodic context, and giving the first higher rank generalizations.

6.1. The source of dynamics. Although the relevance of dynamics to integral points on the sphere is not immediately apparent, it is not difficult to see from an adelic perspective. We have already mentioned in Section 2.2.1 that all three theorems (Theorems 1–3) may be considered as questions about the distribution of an orbit of an adelic torus $z_d.\boldsymbol{T}_d(\boldsymbol{A})$ inside $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$.

One can, therefore, hope to use results about the dynamics of a local torus $\boldsymbol{T}_d(\mathbb{Q}_v)$ acting on $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$ for some fixed place v. However, this is possible only if $\boldsymbol{T}_d(\mathbb{Q}_v)$ noncompact; for otherwise there is no dynamics of interest. This leads to Linnik's condition (cf. Theorem 4), because $\boldsymbol{T}_d(\mathbb{Q}_p)$ is noncompact precisely when $\mathbb{Q}(\sqrt{d})$ is split at p.

6.2. Linnik's method in the light of modern ergodic theory. Much of this joint work is based on the recent work of Einsiedler and Lindenstrauss on classification of invariant measures for toric actions, which is discussed in their contribution to these proceedings [28].

A central concept here is that of *entropy*; we briefly reprise the definition. We recall that if $\mathcal{P}$ is a partition of the probability space (X, ν), the entropy of $\mathcal{P}$ is defined as $h_\nu(\mathcal{P}) := \sum_{S\in\mathcal{P}} -\nu(S)\log\nu(S)$. If T is a measure-preserving transformation of (X, ν), then the measure entropy of T is defined as

$$h(T) = \sup_{\mathcal{P}} \lim_{n\to\infty} \frac{h_\nu(\mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-(n-1)}\mathcal{P})}{n} \tag{6.1}$$

where the supremum is taken over all finite partitions of X.

Here are two results that illustrate the importance of this concept (we denote by Haar the G-invariant probability measure on a quotient space $\Gamma\backslash G$).

The first one is a specialization of the fact that on the unit tangent bundle of a surface of constant negative curvature, the Liouville measure is the unique measure of maximal entropy w.r.t. the action of the geodesic flow:

Fact 1. *Let μ on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ be invariant by the diagonal subgroup, and let a be a nontrivial diagonal matrix. Then $h_\mu(a) \leqslant h_{\mathrm{Haar}}(a)$, with equality if and only if $\mu = \mathrm{Haar}$.*

The second fact lies much deeper and is a result of Einsiedler, Katok and Lindenstrauss [27] which illustrate the phenomenon of *measure rigidity* for the action of tori of rank $\geqslant 2$:

Fact 2. *Let μ be a probability measure on $\mathrm{SL}_3(\mathbb{Z})\backslash\mathrm{SL}_3(\mathbb{R})$ invariant by the diagonal subgroup A and let $a \in A$ be nontrivial. If $h_\mu(a) > 0$ and μ is ergodic (w.r.t. A), then $\mu = \mathrm{Haar}$.*

The scheme of [29, I, II and III] is to treat Linnik problems by combining results of the above type – towards the classification of measures with positive entropy – with diophantine ideas that establish positive entropy. In the subsequent sections we discuss some applications of this general scheme; we have aimed for concreteness, but these methods are much more generally applicable.

6.3. Entropy and the "Linnik principle". In [29, II] we give a purely dynamical proof of Theorem 3. This proof is still based heavily on Linnik's ideas but it introduces considerable conceptual simplification using the notion of entropy discussed in the previous section, and uses in particular *Fact 1*.

We insist that our proof requires *no* splitting condition at some fixed prime p: the reason is that in the context of Theorem 3, the place $v = \infty$ splits in the real quadratic field $\mathbb{Q}(\sqrt{d})$ and so Linnik's condition is *satisfied*! Curiously this was apparently never remarked by Linnik and Skubenko who only used the action of a p-adic split torus.

Let $d > 0$ be a fundamental discriminant. The unit tangent bundle of $Y_0(1)$ is identified with $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$, and so the subset Γ_d described in Theorem 3 may be regarded as a subset $\Gamma_d \subset \mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$. Considered in this way, Γ_d is invariant by the subgroup

$$A = \left\{ a(t) = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix},\ t \in \mathbb{R} \right\}$$

of diagonal matrices with positive entries. It supports a natural A-invariant probability μ_d (the one which assigns the same mass to each connected component) and Theorem 3 asserts precisely that μ_d converge weakly to the $\mathrm{PGL}_2(\mathbb{R})$-invariant probability measure on $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$.

The dynamical proof uses Fact 1 together with a Diophantine computation to show that any weak limit of the μ_d has maximal entropy w.r.t. the action of $a(1)$. The Diophantine computation is a version of "Linnik's basic Lemma," [57, Theorem III.2.1] which in turn may be deduced from *Siegel's mass formula*.

6.4. A rank 3 version of Duke's theorem. A natural "rank 2" version of Theorem 3 is to consider the distribution properties of appropriate collections of compact *flats* inside the Riemannian manifold $\mathrm{PGL}_3(\mathbb{Z})\backslash\mathrm{PGL}_3(\mathbb{R})/\mathrm{PO}_3(\mathbb{R})$.

More generally, let D be a $\mathbb{R}$-split central simple algebra of rank 3 over $\mathbb{Q}$, i.e. $\dim_{\mathbb{Q}} \mathrm{D} = 9$, so that $\mathrm{D} \otimes_{\mathbb{Q}} \mathbb{R} = M_3(\mathbb{R})$. Let $\mathcal{O}_{\mathrm{D}}$ be a fixed maximal order in D. Let $\boldsymbol{G}$ be the algebraic group $\mathrm{PG}(\mathrm{D}) = \mathrm{D}^\times / Z(\mathrm{D})^\times$; we fix a maximal split torus $A = (\mathbb{R}^\times)^2$ inside $\boldsymbol{G}(\mathbb{R})$. Let U be the standard maximal compact subgroup $\prod_p \mathrm{PG}(\mathcal{O}_{\mathrm{D},p})$ of $\boldsymbol{G}(\boldsymbol{A}_f)$. We will assume, for simplicity, that the class number of $\mathcal{O}_{\mathrm{D}}$ is 1, i.e. that $\boldsymbol{G}(\boldsymbol{A}_f) = \boldsymbol{G}(\mathbb{Q}).U$.

Let $K \subset \mathrm{D}$ be a totally real cubic field, together with an isomorphism $\theta\colon K \otimes \mathbb{R} \to \mathbb{R}^3$. We assume for simplicity that $K \cap \mathcal{O}_{\mathrm{D}}$ is the maximal order $\mathcal{O}_K$ of K. This yields, in particular, an embedding of the torus $\boldsymbol{T}_K = \mathrm{res}_{K/\mathbb{Q}} K^\times / \mathbb{Q}^\times$

into the algebraic group PG(D). The choice of θ determines a unique $g_\theta \in \boldsymbol{G}(\mathbb{R})$ so that $g_\theta A g_\theta^{-1} = \boldsymbol{T}_K(\mathbb{R})$. Setting $U_T = \boldsymbol{T}_K(\boldsymbol{A}_f) \cap U$, we consider

$$\Gamma_K := (\boldsymbol{T}_K(\mathbb{Q})\backslash \boldsymbol{T}_K(\boldsymbol{A})/U_T) g_\theta \subset \boldsymbol{G}(\mathbb{Q})\backslash \boldsymbol{G}(\boldsymbol{A})/U \cong \mathcal{O}_{\mathrm{D}}^\times \backslash \mathrm{PGL}_3(\mathbb{R}).$$

This is a collection of compact A-orbits on $\mathcal{O}_{\mathrm{D}}^\times \backslash \mathrm{PGL}_3(\mathbb{R})$, which are indexed by $\mathbf{T}_K(\mathbb{Q})\backslash \mathbf{T}_K(\boldsymbol{A}_f)/U_T$ and the latter quotient is precisely the class group $\mathrm{Cl}(\mathcal{O}_K)$. Consequently, to any subset $S \subset \mathrm{Cl}(\mathcal{O}_K)$ of the class group, we may associate a collection $\Gamma_{K,S}$ of $|S|$ closed A-orbits on $\mathcal{O}_{\mathrm{D}}^\times \backslash \mathrm{PGL}_3(\mathbb{R})$. This set supports a natural A-invariant probability measure (which assigns the same mass to each of the constituent orbits); call this measure $\mu_{K,S}$. For $S = \mathrm{Cl}(\mathcal{O}_K)$ we write simply Γ_K and μ_K.

In [29, I and III] we investigate weak limits of such measures as $\mathrm{disc}(K) \to +\infty$. In the split case, we obtain equidistribution for the full packet of compact orbits which represents the 3-dimensional analog of Theorem 3:

Theorem 10. *Suppose* D *is split* (*i.e.* $\mathrm{D} = M_3(\mathbb{Q})$). *As* $\mathrm{disc}(K) \to \infty$, Γ_K *becomes equidistributed on* $\mathrm{PGL}_3(\mathbb{Z})\backslash \mathrm{PGL}_3(\mathbb{R})$ *with respect to the Haar measure.*

In the non-split case, we obtain a weak form of equidistribution but valid for rather small packets of compact orbits:

Theorem 11. *Suppose* D *is not* $\mathbb{Q}$*-split. Fix* $\delta < 1/2$. *There is a constant* $c = c(\delta) > 0$ *such that, if each set* $S \subset \mathrm{Cl}(\mathcal{O}_K)$ *satisfies* $\frac{|S|}{|\mathrm{Cl}(\mathcal{O}_K)|} \geqslant \mathrm{disc}(K)^{-\delta}$, *then any weak limit of* $\mu_{K,S}$ *contains a Haar component of size* $\geqslant c(\delta)$.

The proof of these results follows the general strategy outlined at the end of Section 6.2 and uses Fact 2 mentioned above as a key ingredient.

In the context of Theorem 10, the first point to verify (since $\mathrm{PGL}_3(\mathbb{Z})\backslash \mathrm{PGL}_3(\mathbb{R})$ is not compact) is that any weak limit of the μ_K, μ say, is a probability measure i.e. that there is no "escape of mass" to ∞. To circumvent this difficulty, we use a version of the harmonic analytic approach described in Section 2: from Mahler's compactness criterion, we build test functions which dominate the characteristic function of small neighborhoods of the cusp and we control escape of mass by bounding the corresponding Weyl sums. The test functions are in fact Eisenstein series (associated with the minimal parabolic of PGL_3) and the resulting Weyl sums can be expressed in terms of the Dedekind zeta function of K, ζ_K to which we apply Corollary 5.2 in the case $F = \mathbb{Q}$. More generally, a variant of this construction together with Corollary 5.2 enable us to bound from above the μ-mass of small neighborhoods of *any* point in $\mathrm{PGL}_3(\mathbb{Z})\backslash \mathrm{PGL}_3(\mathbb{R})$. This is sufficient to imply that *every* ergodic component of μ has positive entropy for some non-trivial $a \in A$. From this we conclude that $\mu = \mu_{\mathrm{Haar}}$ using Fact 2. It should be remarked that the very existence of such test functions is a special feature of the split case.

In the non-split case, there is no issue about possible escape of mass since $\mathcal{O}_{\mathrm{D}}^\times \backslash \mathrm{PGL}_3(\mathbb{R})$ is compact; on the other hand, simple test functions for which the Weyl sums could be evaluated and from which positive entropy could be deduced do

not seem to be available in the cocompact case; instead, we rely on a weaker version of Linnik's basic lemma – *Linnik's principle* – from which we deduce, at least, that a *positive proportion* of the ergodic components of μ have positive entropy; again we conclude by applying Fact 2 to these components. Although it does not achieve equidistribution, Theorem 11 nevertheless illustrates a major advantage of the ergodic approaches of [29] over harmonic-analysis methods: ergodic methods allow for nontrivial results even for *very small* torus orbits ("supersparse equidistribution"): indeed, any exponent $\delta < 1/2$ is admissible and since the size of the class group of $\mathcal{O}_K$ is at most $\mathrm{disc}(K)^{1/2+\varepsilon}$, this is as strong as could be hoped for. Distribution problems for *small* torus orbits arise naturally in several arithmetic questions: for instance we expect that measure rigidity results for actions of p-adic tori should allow for partial progress towards Zhang's measure-theoretic refinement of the André/Oort conjecture [82].

6.5. An application to Minkowski's Theorem. We first recall

Theorem (Minkowski). *Let K be a number field of degree d and maximal order $\mathcal{O}_K$; any ideal class for $\mathcal{O}_K$ possesses an integral representative $J \subset \mathcal{O}_K$ of norm $N(J) = O(\sqrt{\mathrm{disc}(K)})$ where the implicit constant depends only on d.*

We conjecture that this is not sharp for totally real number fields of degree $d \geqslant 3$:

Conjecture 2. Suppose $d \geqslant 3$ is fixed. Then any ideal class in a totally real number fields of degree d has an integral representative of norm $o(\sqrt{\mathrm{disc}(K)})$.

Let $m(K)$ denote the maximum, over ideal classes of $\mathcal{O}_K$, of the minimal norm of a representative. Conjecture 2 asserts that

$$\lim_{\mathrm{disc}(K)\to\infty} \frac{m(K)}{\mathrm{disc}(K)^{1/2}} = 0,$$

for K varying through totally real fields of fixed degree $d \geqslant 3$. It may be shown that for any $d \geqslant 2$ there exists a $c' > 0$ such that there is an infinite set of totally real fields of degree d for which $m(K) \geqslant c' \cdot \mathrm{disc}(K)^{1/2}(\log \mathrm{disc}(K))^{1-2d}$. Thus Minkowski's Theorem is rather close to sharp and in fact Conjecture 2 is unlikely to be true for $d = 2$. For $d \geqslant 3$ this conjecture can be seen as a result of the extra freedom that arises from having a group of units of rank $d - 1 \geqslant 2$, and is actually a consequence of a stronger conjecture formulated by Margulis [59].

We will call an ideal class of a field K δ-bad if it does not admit a representative of norm $< \delta \cdot \mathrm{disc}(K)^{1/2}$. Let $h_\delta(K)$ be the number of δ-bad ideal classes and let R_K denote the regulator of the field K. In [29, I] it is shown that:

Theorem 12. *Let $d \geqslant 3$, and let K denote a totally real number field of degree d. For all $\varepsilon, \delta > 0$ we have*

$$\sum_{\mathrm{disc}(K)<X} R_K h_\delta(K) \ll X^\varepsilon. \tag{6.2}$$

In particular:

1. *"Conjecture* 2 *is true for almost all fields": the number of fields K with discriminant $\leqslant X$ for which $m(K) \geqslant \delta \cdot \mathrm{disc}(K)^{1/2}$ is $O_{\varepsilon,\delta}(X^{\varepsilon})$, for any $\varepsilon, \delta > 0$.*

2. *"Conjecture* 2 *is true for all fields with large regulator": If $(K_i)_i$ is any sequence of fields for which $\liminf_i \frac{\log R_{K_i}}{\log \mathrm{disc}(K_i)} > 0$, then $m(K_i) = o(\mathrm{disc}(K_i)^{1/2})$.*

This is connected to the considerations of Section 6.4 in the following way. Consider the case $d = 3$; to a real cubic field K and suitable additional data we have associated a collection of compact A-orbits $\Gamma_K \subset \mathrm{PGL}_3(\mathbb{Z})\backslash\mathrm{PGL}_3(\mathbb{R})$, indexed by the class group of K. The key point is the following: the question of the minimal norm of a representative for a given ideal class is closely related to the question of how far the associated A-orbit penetrates into the "cusp" of the noncompact space $\mathrm{PGL}_3(\mathbb{Z})\backslash\mathrm{PGL}_3(\mathbb{R})$. This allows a geometric reformulation of Theorem 12 that is amenable to analysis by the methods of Section 6.4.

7. Ergodic theory vs. harmonic analysis

In this concluding section, we briefly compare dynamical methods and harmonic analysis.

Fundamentally, the most general type of problem we are considering is the following: let $\boldsymbol{H} \subset \boldsymbol{G}$ be a subgroup of a semisimple $\mathbb{Q}$-group $\boldsymbol{G}$; understand the "distribution" of $\boldsymbol{H}(\mathbb{Q})\backslash\boldsymbol{H}(\boldsymbol{A})$ inside $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$. Indeed such problems arise naturally in a large number of arithmetic questions. Two possible approaches to these questions are the following:

1. Ergodic. Here we choose a suitable finite set of places S and apply results constraining $\boldsymbol{H}(\mathbb{Q}_S)$-invariant measures on $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$.

2. Harmonic-analytic. Here we choose a suitable basis φ_i for functions on $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$ and compute the "periods"

$$\int_{\boldsymbol{H}(\mathbb{Q})\backslash\boldsymbol{H}(\boldsymbol{A})} \varphi_i, \tag{7.1}$$

   the main goal being to have "good" quantitative upper bound for (7.1)

Moreover, we note that there is considerable potential for interaction between the two approaches: in [29, I], harmonic analysis is used to control escape of mass issues, while in [75] quantitative ideas from ergodic theory are used to give estimates on periods like (7.1).

In any case, the following general principles tend to apply:

1. If $\boldsymbol{H}$ is "a large enough subgroup" of $\boldsymbol{G}$ (say if $\boldsymbol{H}$ acts with an open orbit on the flag variety of $\boldsymbol{G}$), the periods (7.1) will often have "arithmetic significance", i.e. are often interpretable in terms of quantities of arithmetic interest such as L-functions and one can at least *hope* for complete, quantitative results via harmonic analysis. Note that, in addition to "standard harmonic analysis," one should keep in mind the possibility of using an extra important trick: namely, of using *equalities between periods on different groups*. That is to say: often there will be another pair $(\boldsymbol{H}' \subset \boldsymbol{G}')$ with the property that, for each φ_i as above, one may associate functions φ_i' on $\boldsymbol{G}'(\mathbb{Q})\backslash\boldsymbol{G}'(\boldsymbol{A})$ so that

$$\int_{\boldsymbol{H}(\mathbb{Q})\backslash\boldsymbol{H}(\boldsymbol{A})} \varphi_j = \int_{\boldsymbol{H}'(\mathbb{Q})\backslash\boldsymbol{H}'(\boldsymbol{A})} \varphi_j'.$$

   The correspondence $\varphi \leftrightarrow \varphi'$ is usually related to functoriality. Thereby one can study the $\boldsymbol{H}$-periods on $\boldsymbol{G}$ by switching to $\boldsymbol{G}'$.

2. If $\boldsymbol{H}$ is not a torus, one often apply profitably Ratner's theorem in the ergodic approach and get strong, *although non-quantitative* results. We have not discussed any examples of this in the present article; a nice instance is [30].

3. If $\boldsymbol{H}$ is a torus, the emerging theory of measure rigidity for torus actions (see in particular [27], [28]) may offer a substitute for Ratner theory. This requires an extra input, positive entropy, and has two further disadvantages (compared to "Ratner theory") that might be noted:

   (a) At present there is no good general way to control, either escape of mass when $\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A})$ is noncompact, or the related phenomenon of concentration on embedded subgroups.

   (b) One needs to have "Linnik's condition," i.e. a fixed set of places S such that $\boldsymbol{H}(\mathbb{Q}_v)$ is noncompact for $v \in S$.

Eventually, as a rough rule, the strength of ergodic theory is that it can handle orbits of very "small" subgroups – which at present seem far beyond the reach of traditional harmonic analysis– and its weakness is that it is not (yet) quantitative. On the other hand, the strength of harmonic analysis is that it imports all the rich internal structure of automorphic forms.

For instance, "why" is it that harmonic-analytic approaches to Theorem 1 have been able to avoid a Linnik-type condition? Our perspective to this question is that the Waldspurger formula (2.5) expresses a period over a non-split torus $\boldsymbol{T}_d$ in terms of the L-function $L(\pi, 1/2) \times L(\pi \times \chi_d, 1/2)$. But, by Hecke theory, the second L-function is expressible as a (χ_d-twisted) period of a form in π over a *split* torus in $\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\boldsymbol{A}_{\mathbb{Q}})$! Thereby one has an *equality* between a period over a *nonsplit* torus $\boldsymbol{T}_d$ and a twisted period over a *split* torus $\boldsymbol{T}_{\mathrm{split}}$. This equality, which is an instance of functoriality, is part of the reason that one is able to sidestep the problem of small split primes that plagues any direct analysis of $\boldsymbol{T}_d$.

References

[1] Baruch, E. M., Mao, Z., Central value of automorphic L-functions. Preprint, 2003.

[2] Bernstein, J., Reznikov, A., Analytic continuation of representations and estimates of automorphic forms. *Ann. of Math.* (2) **150** (1) (1999), 329–352.

[3] —, Estimates of automorphic functions. *Mosc. Math. J.* **(4)** (1) (2004), 19–37, 310

[4] —, Periods, subconvexity and representation theory. *J. Differential. Geometry* **70** (1) (2005), 129–141.

[5] Blomer, V., Non-vanishing of class group L-functions at the central point. *Ann. Inst. Fourier* (*Grenoble*) **54** (4) (2004), 831–847.

[6] —, Shifted convolution sums and subconvexity bounds for automorphic L-functions. *Internat. Math. Res. Notices* (2004), (73), 3905–3926.

[7] Blomer, V., Harcos,G., Michel, Ph., A Burgess-like subconvex bound for twisted L-functions. *Forum Math.* (2006), to appear.

[8] —, Bounds for automorphic L-functions. Preprint, 2006.

[9] Bykovskiĭ, V. A., A trace formula for the scalar product of Hecke series and its applications. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **226**, (1996), 14–36, 235–236; English transl. *J. Math. Sci.* (*New York*) **89** (1) (1998), 915–932.

[10] Clozel, L., Oh, H., Ullmo, E., Hecke operators and equidistribution of Hecke points. *Invent. Math.* **144** (2) (2001), 327–351.

[11] Clozel, L., Ullmo, E., Équidistribution des points de Hecke. In *Contributions to automorphic forms, geometry, and number theory*, Johns Hopkins University Press, Baltimore, MD, 2004, 193–254.

[12] —, Équidistribution de mesures algébriques, *Compositio Math.* **141** (5) (2005), 1255–1309.

[13] Cogdell, J. W., On sums of three squares. *J. Théor. Nombres Bordeaux* **15** (1) (2003), 33–44.

[14] Cohen, P. B., Hyperbolic equidistribution problems on Siegel 3-folds and Hilbert modular varieties. *Duke Math. J.* **129** (1) (2005), 87–127.

[15] Conrey, J. B., Iwaniec, H., The cubic moment of central values of automorphic L-functions. *Ann. of Math.* (2) **151** (3) (2000), 1175–1216.

[16] Cornut, C., Vatsal, V., Nontriviality of Rankin-Selberg L-functions and CM points. In *Proceedings of the LMS symposium: L-functions and Galois representations* (2004), to appear.

[17] Duke, W., Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.* **92** (1) (1988), 73–90.

[18] Duke, W., Friedlander, J. B., Iwaniec, H., A quadratic divisor problem. *Invent. Math.* **115** (2) (1994), 209–217.

[19] —, Bounds for automorphic L-functions. *Invent. Math.* **112** (1) (1993), 1–8.

[20] —, Bounds for automorphic L-functions. II. *Invent. Math.* **115** (2) (1994), 219–239.

[21] —, Bilinear forms with Kloosterman fractions. *Invent. Math.* **128** (1) (1997), 23–43.

[22] —, Bounds for automorphic L-functions. III. *Invent. Math.* **143** (2) (2001), 221–248.

[23] —, The subconvexity problem for Artin L-functions. *Invent. Math.* **149** (3) (2002), 489–577.

[24] —, Class group L-functions. *Duke Math. J.* **79** (1) (1995), 1–56.

[25] Duke, W., Rudnick, Z., Sarnak, P., Density of integer points on affine homogeneous varieties, *Duke Math. J.* **71** (1) (1993), 143–179.

[26] Duke, W., Schulze-Pillot, R., Representation of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids. *Invent. Math.* **99** (1) (1990), 49–57.

[27] Einsiedler, M., Katok, A., Lindenstrauss, E., Invariant measures and the set of exceptions to Littlewoods conjecture. *Ann. of. Math.* (2006), to appear.

[28] Einsiedler, M., Lindenstrauss, E., Diagonalizable flows on locally homogeneous spaces and number theory. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1731–1759.

[29] Einsiedler, M., Lindenstrauss, E., Michel, Ph., Venkatesh, A., Distribution properties of compact orbits on homonegeous spaces I, II & III. In preparation.

[30] Eskin, A., Oh, H., Representations of integers by an invariant polynomial and unipotent flows. Preprint, 2003.

[31] Eskin, A., McMullen, C., Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.* **71** (1) (1993), 181–209.

[32] Eskin, A., Mozes, Sh., Shah, N., Unipotent flows and counting lattice points on homogeneous varieties. *Ann. of Math.* (2) **143** (2) (1996), 253–299.

[33] Eskin, A., Counting problems and semisimple groups. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 539–552.

[34] Friedlander, J. B., Bounds for L-functions. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 363–373.

[35] Gan, W. T., Oh, H., Equidistribution of integer points on a family of homogeneous varieties: a problem of Linnik. *Compositio Math.* **136** (3) (2003), 323–352.

[36] Gross, B., Heights and the special values of L-series. In *Number theory* (Montreal, Que., 1985), CMS Conf. Proc. 7, Amer. Math. Soc., Providence, RI, (1987, 115–187.

[37] Gross, B., Zagier, D., Heegner points and derivatives of L-series. *Invent. Math.* **84** (2) (1986), 225–320.

[38] Harcos, G., An additive problem in the Fourier coefficients of cusp forms. *Math. Ann.* **326** (2) (2003), 347–365.

[39] Harcos, G., Michel, Ph., The subconvexity problem for Rankin-Selberg L-functions and equidistribution of Heegner points II. *Invent. Math.* **163** (3) (2006), 581–655.

[40] Harris, M., Kudla, S. S., The central critical value of a triple product L-function. *Ann. of Math.* (2) **133** (3) (1991) 605–672.

[41] Iwaniec, H., Fourier coefficients of modular forms of half-integral weight. *Invent. Math.* **87** (2) (1987), 385–401.

[42] —, The spectral growth of automorphic L-functions. *J. Reine Angew. Math.* **428** (1992), 139–159.

[43] —, Harmonic analysis in number theory. In *Prospects in mathematics* (Princeton, NJ, 1996), Amer. Math. Soc., Providence, RI, 1999, 51–68.

[44] Iwaniec, H., Sarnak, P., Perspectives on the analytic theory of L-functions. *Geom. Funct. Anal.* (2000) Special Volume, 705–741.

[45] Jutila, M., Convolutions of Fourier coefficients of cusp forms. *Publ. Inst. Math.* (*Beograd*) (*N.S.*) **65 (79)** (1999), 31–51.

[46] Jutila, M., Motohashi, Y., Uniform bounds for Hecke L-functions. *Acta Math.* **195** (2005), 61–115.

[47] Katok, S., Sarnak, P., Heegner points, cycles and Maass forms. *Israel J. Math.* **84** (1–2) (1993), 193–227.

[48] Kim, H., Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. *J. Amer. Math. Soc.* **16** (1) (2003), 139–183.

[49] Kim, Henry H., Shahidi, Freydoon, Functorial products for $\mathrm{GL}_2 \times \mathrm{GL}_3$ and the symmetric cube for GL_2. *Ann. of Math.* (2) **155** (3) (2002), 837–893.

[50] Kohnen, W., Zagier, D., Values of L-series of modular forms at the center of the critical strip. *Invent. Math.* **64** (2) (1981), 175–198.

[51] Kowalski, E., Michel, Ph., VanderKam, J., Mollification of the fourth moment of automorphic L-functions and arithmetic applications. *Invent. Math.* **142** (1) (2000), 95–151.

[52] —, Rankin-Selberg L-functions in the level aspect. *Duke Math. J.* **114** (1) (2002), 123–191.

[53] Lindenstrauss, E., Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math.* (2) **163** (1) (2006), 165–219.

[54] Linnik, Yu. V., The asymptotic distribution of reduced binary quadratic forms in relation to the geometries of Lobačevskiĭ. III. *Vestnik Leningrad. Univ.* **10** (8) (1955), 15–27.

[55] —, Asymptotic-geometric and ergodic properties of sets of lattice points on a sphere. *Amer. Math. Soc. Transl.* (2) **13** (1960), 9–27.

[56] —, Additive problems and eigenvalues of the modular operators. In *Proceedings of the International Congress of Mathematicians* (Stockholm, 1962), Inst. Mittag-Leffler, Djursholm 1963, 270–284.

[57] —, *Ergodic properties of algebraic fields*. Ergeb. Math. Grenzgeb. 45, Springer-Verlag, New York 1968.

[58] Linnik, Yu. V., Skubenko, B. F., Asymptotic distribution of integral matrices of third order. *Vestnik Leningrad. Univ. Ser. Mat. Meh. Astronom.* **19** (3) (1964), 25–36.

[59] Margulis, G., Problems and conjectures in rigidity theory. In *Mathematics: frontiers and perspectives*, Amer. Math. Soc., Providence, RI, 2000, 161–174.

[60] Michel, Ph., The subconvexity problem for Rankin-Selberg L-functions and equidistribution of Heegner points. *Ann. of Math.* (2), **160** (1) (2004), 185–236.

[61] Michel, Ph., Venkatesh, A., Periods, subconvexity and equidistribution. In preparation.

[62] —, Heegner points and nonvanishing of Rankin-Selberg L-functions. Preprint, 2006.

[63] Molteni, G., Upper and lower bounds at $s = 1$ for certain Dirichlet series with Euler product. *Duke Math. J.* **111** (1) (2002), 133–158.

[64] Motohashi, Y., *Spectral theory of the Riemann zeta-function*. Cambridge Tracts in Math., Cambridge University Press, Cambridge 1997.

[65] Oh, Hee, Hardy-Littlewood system and representations of integers by an invariant polynomial. *Geom. Funct. Anal.* **14** (4) (2004), 791–809.

[66] Popa, A., Central values of Rankin L-series over real quadratic fields. *Compositio Math.* (2006), to appear.

[67] Reznikov, A., Rankin-Selberg without unfolding. Preprint, 2005.

[68] Sarnak, P., Diophantine problems and linear groups. In *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. I, The Mathematical Society of Japan, Tokyo, Springer-Verlag, Tokyo, 1991, 459–471.

[69] —, Integrals of products of eigenfunctions. *Internat. Math. Res. Notices* **1994** (6) (1994), 251–260.

[70] —, Estimates for Rankin-Selberg L-functions and quantum unique ergodicity. *J. Funct. Anal.* **184** (2) (2001)n 419–453.

[71] Skubenko, B. F., The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems. *Izv. Akad. Nauk SSSR Ser. Mat.* **26** (1962), 721–752.

[72] Ullmo, E., Théorie ergodique et géométrie arithmétique, In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 197–206.

[73] Vatsal, V., Uniform distribution of Heegner points. *Invent. Math.* **148** (1) (2002), 1–46.

[74] —, Special values of anticyclotomic L-functions. *Duke Math. J.* **116** (2) (2003), 219–261.

[75] Venkatesh, A., Sparse equidistribution problems, period bounds, and subconvexity. Preprint, 2005.

[76] Waldspurger, J.-L., Sur les valeurs de certaines fonctions L automorphes en leur centre de symétrie. *Compositio Math.* **54** (2) (1985), 173–242.

[77] Watson, T., Rankin triple products and quantum chaos, *Ann. of Math.* (2006), to appear.

[78] Xue, H., Central values of Rankin L-functions. Preprint, 2005.

[79] Yafaev, A., A conjecture of Yves André's. *Duke Math. J.* **132** (3) (2006), 393–407.

[80] Zhang, S., Heights of Heegner points on Shimura curves. *Ann. of Math.* (2) **153** (1) (2001), 27–147.

[81] —, Gross-Zagier formula for GL_2. *Asian J. Math.* **5** (2) (2001) 183–290.

[82] —, Equidistribution of CM-points on quaternion Shimura varieties. Preprint, 2004.

Département de Mathématiques, Université Montpellier II, Place E. Bataillon,
34095 Montpellier cedex, France

Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, U.S.A.

p-adic motivic cohomology in arithmetic

Wiesława Nizioł

Abstract. We will discuss recent progress by many people in the program of representing motivic cohomology with torsion coefficients of arithmetic schemes by various arithmetic p-adic cohomologies: étale, logarithmic de Rham–Witt, and syntomic. To illustrate possible applications in arithmetic geometry we will sketch proofs of the absolute purity conjecture in étale cohomology and comparison theorems of p-adic Hodge theory.

Mathematics Subject Classification (2000). Primary 14F42, 14F20, 11G25.

Keywords. motivic cohomology, K-theory, étale cohomology, crystalline cohomology, p-adic Hodge theory.

1. Introduction

In this report, we will survey results about the relationship between motivic cohomology and p-adic cohomologies of arithmetic schemes. We will concentrate on the stable ranges where the motivic cohomology with torsion coefficients tends to be represented by various cohomologies of arithmetic type. Much has happened in that subject in the last few years and a detailed survey can be found in [10]. For us motivic cohomology will mean Bloch higher Chow groups and its approximation – the gamma gradings of algebraic K-theory. They are defined using algebraic cycles and vector bundles, respectively, and are connected via an Atiyah–Hirzebruch spectral sequence. By inverting the Bott element or in some stable ranges motivic cohomology becomes isomorphic to its (étale) topological version. That is the content of the Beilinson–Lichtenbaum and Quillen–Lichtenbaum conjectures both of which follow (over fields) from the Bloch–Kato conjecture (now proved at least for mod 2 coefficients). Torsion étale motivic cohomology has a direct relationship with various arithmetic cohomologies: logarithmic de Rham–Witt (an arithmetic version of crystalline cohomology in positive characteristic), syntomic cohomology (an arithmetic version of crystalline cohomology in mixed characteristic), and arithmetic étale cohomology. As a result we can represent p-adic cohomology classes as algebraic cycle classes in a way reminiscent of the classical situation. This turns out to be – both conceptually and technically – a powerful tool. We present two applications: a proof of the absolute purity conjecture in étale cohomology and proofs of comparison theorems in p-adic Hodge theory.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. K-theory

2.1. Milnor K-theory. For a field k, the Milnor K-groups $K_*^M(k)$ are defined to be the quotient of the tensor algebra of the multiplicative group k^* of k by the ideal generated by the Steinberg relation $x \otimes (1-x)$ for $x \neq 0, 1$. This rather simple construction turns out to be a fundamental object in the subject.

If m is relatively prime to the characteristic of k, Kummer theory gives that $K_1^M(k)/m \simeq k^*/k^{*m} \xrightarrow{\sim} H^1(k_{\text{ét}}, \mu_m)$. Using cup product on Galois cohomology we get the Galois symbol map

$$K_n^M(k)/m \to H^n(k_{\text{ét}}, \mu_m^{\otimes n}).$$

Conjecture 2.1 (Bloch–Kato). The above symbol map is an isomorphism.

Voevodsky proved [40] the Bloch–Kato conjecture for m a power of 2 (Milnor conjecture) and has recently announced a proof for general m [41].

If $p > 0$ is equal to the characteristic of k, the Bloch–Kato conjecture could be interpreted as the theorem of Bloch–Gabber–Kato [3] giving an isomorphism

$$\operatorname{dlog}\colon K_n^M(k)/p^r \xrightarrow{\sim} H^0(k_{\text{ét}}, \nu_r^n),$$

where $\nu_r^n = W_r\Omega^n_{X,\log}$ is the logarithmic de Rham–Witt sheaf. It is a subsheaf of the de Rham–Witt sheaf $W_r\Omega_X^n$ generated locally for the étale topology by $\operatorname{dlog}\bar{x}_1 \wedge \cdots \wedge \operatorname{dlog}\bar{x}_n$, where $\bar{x} \in W_r\mathcal{O}_X$ are the Teichmüller lifts of units. It fits into the short exact sequence of pro-sheaves (F is the Frobenius)

$$0 \to \nu_\cdot^n \to W_\cdot\Omega_X^n \xrightarrow{F-1} W_\cdot\Omega_X^n \to 0.$$

2.2. Algebraic K-theory. For a noetherian scheme X, let $\mathcal{M}(X)$ and $\mathcal{P}(X)$ denote the categories of coherent and locally free sheaves on X, respectively. The higher algebraic K and K' groups of X are defined as homotopy groups of certain simplicial spaces associated to the above categories: $K_i(X) = \pi_i(\mathcal{K}(X))$, $K'_i(X) = \pi_i(\mathcal{K}'(X))$. For $i = 0$, $K_0(X)$ and $K'_0(X)$ are the Grothendieck groups of vector bundles and coherent sheaves, respectively. For a field k, the product structure on K-theory gives a natural homomorphism $K_n^M(k) \to K_n(k)$ that is an isomorphism for $n \leq 2$.

K-groups mod m, $K_i(X, \mathbb{Z}/m)$, are defined by taking homotopy groups with $\mathbb{Z}/m$ coefficients of the above spaces. They are related to $K_*(X)$ by a universal coefficient sequence. Exterior powers of vector bundles induce a descending filtration $F^*_\gamma K_*(X, \mathbb{Z}/m)$ (γ-filtration) and the graded pieces $\operatorname{gr}^*_\gamma K_*(X, \mathbb{Z}/m)$ can be considered an approximation to motivic cohomology groups.

Similarly, we get groups $K'_i(X, \mathbb{Z}/m)$. The natural homomorphism

$$K_i(X, \mathbb{Z}/m) \to K'_i(X, \mathbb{Z}/m)$$

is an isomorphism if X is regular ("Poincaré duality"), because every coherent sheaf has a finite resolution by locally free sheaves. For Z a closed subscheme of X with open complement U, there is a localization sequence in K'-theory

$$\to K'_{i+1}(U, \mathbb{Z}/m) \to K'_i(Z, \mathbb{Z}/m) \to K'_i(X, \mathbb{Z}/m) \to K'_i(U, \mathbb{Z}/m) \to$$

It is relatively easy to derive in this generality and (in arithmetic applications) it gives K-theory great technical advantage over other motivic cohomologies. The other important technical advantage is the ease with which one can define higher Chern classes into various cohomologies. To be able to follow Gillet's construction [17] all one really needs is that the cohomology of the classifying simplicial scheme BGL (and some of its variants) is the expected one.

Varying X, one can view $\mathcal{K}(X)$ as a presheaf of simplicial spaces. Let $\mathcal{K}/m$ denote the presheaf of corresponding spaces mod m. Assume that X is regular. Then the Mayer–Vietoris property of K-theory gives that $K_*(X, \mathbb{Z}/m) \simeq H^{-*}(X_{\mathrm{Zar}}, \mathcal{K}/m)$.

If m is invertible on X, under some additional technical assumptions on X, the étale K-theory of Dwyer–Friedlander [5] can be computed using presheaves $\mathcal{K}/m$: $K_j^{\text{ét}}(X, \mathbb{Z}/m) \simeq H^{-j}(X_{\text{ét}}, \mathcal{K}/m)$, for $j \geq 0$.

Conjecture 2.2 (Quillen–Lichtenbaum). The change of topology map

$$\rho_j \colon K_j(X, \mathbb{Z}/m) \to K_j^{\text{ét}}(X, \mathbb{Z}/m)$$

is an isomorphism for $j \geq \mathrm{cd}_m X_{\text{ét}}$ (the étale cohomological dimension of X).

Here and below we will review the current status of the Quillen–Lichtenbaum conjecture. Recall that Thomason [37] proved that the map ρ_j induces an isomorphism

$$\tilde{\rho}_j \colon K_j(X, \mathbb{Z}/m)[\beta_m^{-1}] \xrightarrow{\sim} K_j^{\text{ét}}(X, \mathbb{Z}/m),$$

where $K_j(X, \mathbb{Z}/m)[\beta_m^{-1}]$ denotes the j'th graded piece of the ring obtained by inverting the action of the Bott element β_m on $K_*(X, \mathbb{Z}/m)$. If $\mu_m \subset \Gamma(X, \mathcal{O}_X)$, then β_m is defined as an element in $K_2(X, \mathbb{Z}/m)$ canonically lifting a chosen primitive m'th root of unity in $K_1(X)$. A refined version of the proof of this theorem [38] allowed him to show that for a variety of schemes (not necessarily over an algebraically closed fields) the map ρ_j is surjective for j larger than (roughly) $N = (\dim X)^3$ and its kernel is annihilated by β_m^N. Over an algebraically closed field and for quasi-projective X we can do better: Walker shows [42] that in that case ρ_j is split surjective for $j \geq 2d$ and its kernel is annihilated by β_m^d, $d = \dim X$ (see also [12]).

Étale K-theory has a direct relationship to étale cohomology. Namely Gabber's rigidity and Suslin's computation of K-theory of algebraically closed fields imply that the sheaves of fundamental groups $\tilde{\pi}_i(\mathcal{K}/m)$ are isomorphic to $\mu_m^{\otimes i/2}$ for i even and are 0 for i odd. Then the local to global spectral sequence becomes

$$E_2^{p,q} = \begin{cases} H^p(X_{\text{ét}}, \mu_m^{q/2}) & \text{for } q \text{ even}, \\ 0 & \text{for } q \text{ odd} \end{cases} \quad \Rightarrow \quad K_{q-p}^{\text{ét}}(X, \mathbb{Z}/m).$$

Action of Adams operations shows that this spectral sequence degenerates at E_2 modulo torsion of a bounded order depending only on $\mathrm{cd}_m X_{\text{ét}}$. In particular, the Chern classes

$$c^{\text{ét}}_{i,j}\colon \operatorname{gr}^i_\gamma K^{\text{ét}}_j(X, \mathbb{Z}/m) \to H^{2i-j}(X_{\text{ét}}, \mu_m^{\otimes i})$$

are isomorphisms modulo small torsion.

If X is smooth over a perfect field of characteristic $p > 0$, Geisser–Levine [15] using motivic cohomology (see below) prove the isomorphism $\tilde{\pi}_n(\mathcal{K}/p^r) \simeq \nu^n_r$. Since ν^n_r vanishes for $n > \dim X$, the local to global spectral sequence

$$H^k(X, \tilde{\pi}_n(\mathcal{K}/p^r)) \Rightarrow K_{n-k}(X, \mathbb{Z}/p^r)$$

gives the important vanishing result: $K_n(X, \mathbb{Z}/p^r) = 0$ for $n > \dim X$.

2.3. Application: the absolute purity conjecture. The relationship between algebraic K-theory and étale cohomology just described was used by Thomason [36] and Gabber [13] to prove the absolute purity conjecture in étale cohomology. Thomason derived it (up to small torsion) from absolute purity in K-theory and Gabber – after some reductions – from vanishing results in K-theory.

Conjecture 2.3. Let $i\colon Y \hookrightarrow X$ be a closed immersion of noetherian, regular schemes of pure codimension d. Let n be an integer invertible on X. Then

$$\mathcal{H}^q_Y(X_{\text{ét}}, \mathbb{Z}/n) \simeq \begin{cases} 0 & \text{for } q \neq 2d, \\ \mathbb{Z}/n(-d) & \text{for } q = 2d. \end{cases}$$

Proof. Thomason's proof works (for example) for schemes of finite type over $\mathbb{Z}$ and m all of whose prime divisors are at least $\dim X+1$. Localization sequence immediately gives absolute purity in K-theory: one defines K-theory with support $K_Y(X)$ to be the homotopy fiber of the restriction $K(X) \to K(X \setminus Y)$ and by localization and Poincaré duality we get the isomorphism $K_{Y,*}(X) \simeq K_*(Y)$. Inverting the Bott element yields absolute cohomological purity in étale K-theory: $K^{\text{ét}}_{Y,*}(X) \simeq K^{\text{ét}}_*(Y)$. This can be now easily transferred to étale cohomology via the Atiyah–Hirzebruch spectral sequence. Namely, we have the sheafification of the Atiyah–Hirzebruch spectral sequence with support

$$E_2^{p,q} = \begin{cases} \mathcal{H}^p_Y(X_{\text{ét}}, \mu_m^{\otimes i}) & \text{for } q = 2i, \\ 0 & \text{for } q \neq 2i \end{cases}$$

strongly converging to the sheaf associated to $K^{\text{ét}}_{Y,q-p}(-, \mathbb{Z}/m) \simeq K^{\text{ét}}_{q-p}(-_Y, \mathbb{Z}/m)$. Evoking once more the Atiyah–Hirzebruch spectral sequence (this time on Y) one computes that this sheaf is periodic (of period two) and $\mathbb{Z}/m$ for $q = p$ and trivial for $q - p = 1$. Action of Adams operations shows that the above spectral sequence degenerates modulo a constant (depending on étale cohomological dimension of X) and that the same constant kills $E_\infty^{2j,2j}$ for $j \neq q$.

To prove absolute purity in general Gabber appeals to the Atiyah–Hirzebruch spectral sequence only in the situation where it does in fact degenerate. First, he defines a well-behaved global cycle class $c(Y) \in H^{2d}_Y(X_{\text{ét}}, \mu_m^{\otimes d})$ that allows him to reduce absolute purity to a punctual one: for a regular strict local ring $\mathcal{O}$ of dimension d with closed point $i_x\colon x \to \operatorname{Spec}\mathcal{O}$ the cycle class gives an isomorphism $cl(x)\colon \mu_{m,x} \simeq i_x^! \mu_m^{\otimes d}[2d]$. Induction now reduces it to a vanishing result: $H^p(\mathcal{O}[f^{-1}]_{\text{ét}}, \mu_m) = 0$ for $p \neq 0, 1$, where $f \in \mathfrak{m} \setminus \mathfrak{m}^2$ for the maximal ideal $\mathfrak{m} \subset \mathcal{O}$. Here he can assume $\mathcal{O}$ to be of arithmetic type.

Next, he considers the following Atiyah–Hirzebruch spectral sequence

$$E_2^{p,q} = \begin{cases} H^p(\mathcal{O}[f^{-1}]_{\text{ét}}, \mu_m^{q/2}) & \text{for } q \text{ even}, \\ 0 & \text{for } q \text{ odd} \end{cases} \quad \Rightarrow K^{\text{ét}}_{q-p}(\mathcal{O}[f^{-1}], \mathbb{Z}/m)$$

where the étale K-groups $K^{\text{ét}}_{q-p}(\mathcal{O}[f^{-1}], \mathbb{Z}/m)$ are equal to $K_0(\mathcal{O}[f^{-1}], \mathbb{Z}/m)(q/2)$ for q even and $K_1(\mathcal{O}[f^{-1}], \mathbb{Z}/m)((q-1)/2)$ for q odd. Inductively, using local affine Lefschetz and duality he gets vanishing of $H^p(\mathcal{O}[f^{-1}]_{\text{ét}}, \mu_m) = 0$ for $p \neq 0, 1, d-1, d$. That kills some colums in the above spectral sequence and the degeneration at E_2 follows. Now, vanishing of level 2 of the filtration on K-groups implies that $E_2^{p,q} = E_\infty^{p,q} = 0$ for $p \geq 2$ yielding $H^p(\mathcal{O}[f^{-1}]_{\text{ét}}, \mu_m) = 0$ for $p \neq 0, 1$, as wanted. □

3. Motivic cohomology

3.1. Motivic cohomology over a field. For a separated scheme X over a field, Bloch higher Chow groups [1] are the cohomology groups of a certain complex of abelian groups. To define this complex, denote by $\triangle^n$ the algebraic n-simplex $\operatorname{Spec}\mathbb{Z}[t_0, \ldots, t_n]/\big(\sum t_i - 1\big)$. Let $z^r(X, i)$ denote the free abelian group generated by irreducible codimension r subvarieties of $X \times \triangle^i$ meeting all faces properly. Let $z^r(X, *)$ be the chain complex thus defined with boundaries given by pullbacks of cycles along face maps. Denote by $H^i(X, \mathbb{Z}(r))$ the cohomology of the complex $\mathbb{Z}(r) := z^r(X, 2r - *)$ in degree i. The motivic cohomology with coefficients $\mathbb{Z}/m$ is the cohomology of the complex $\mathbb{Z}/m(r) = \mathbb{Z}(r) \otimes \mathbb{Z}/m$. It fits into the usual universal coefficient sequence.

Remark 3.1. There is another commonly used construction of motivic cohomology due to Suslin–Voevodsky [9]. It works well in characteristic zero but it is not well suited for studying mod p-phenomena in positive characteristic p.

Motivic cohomology groups are trivial for $i > 2n$ and $i > n + \dim X$ for dimension reasons. The Beilinson–Soulé conjecture (still open) postulates that they vanish for $i < 0$. We have $H^{2n}(X, \mathbb{Z}(n)) \simeq CH_n(X)$, the classical Chow group. For a field k, $H^i(k, \mathbb{Z}(n))$ are trivial for $i > n$ and agree with the Milnor group $K_n^M(k)$ for $i = n$. In particular, $H^n(k, \mathbb{Z}/m(n)) \simeq K_n^M(k)/m$.

For Z a closed subscheme of X of codimension c with open complement U, there is a localization sequence

$$\to H^{i-2c}(Z, \mathbb{Z}(n-c)) \to H^i(X, \mathbb{Z}(n)) \to H^i(U, \mathbb{Z}(n)) \to H^{i+1-2c}(Z, \mathbb{Z}(n-c)) \to$$

It is a difficult theorem to prove. It is also rather difficult in general to construct higher cycle classes $c_{i,n}\colon H^i(X, \mathbb{Z}(n)) \to H^i(X, n)$ into various bigraded cohomology theories relevant to arithmetic. The original method of Bloch [2] requires weak purity as well as homotopy property both of which fail for some commonly used p-adic cohomologies. In particular, we are still missing a definition of cycle classes into syntomic cohomology independent of the theory of p-adic periods.

It is even more difficult to show that the higher Chow groups and K'-theory are related by an Atiyah–Hirzebruch spectral sequence

$$E_2^{s,t} = H^{s-t}(X, \mathbb{Z}(-t)) \Rightarrow K'_{-s-t}(X). \tag{3.1}$$

This sequence was first constructed for fields by Bloch–Lichtenbaum [4], then generalized to quasi-projective varieties by Friedlander–Suslin [11], and finally to schemes of finite type by Levine [25],[23]. By different methods, it was also constructed by Grayson–Suslin [18], [34] and Levine [26]. If X is regular, the action of Adams operations shows that this sequence degenerates modulo small torsion and the resulting filtration differs from the γ-filtration by a small torsion. In particular, we get that

$$\mathrm{gr}^i_\gamma K_j(X) \otimes \mathbb{Q} \simeq H^{2i-j}(X, \mathbb{Q}(i)).$$

Varying X, one gets a sheaf $\mathbb{Z}(n) := z^n(-, 2n - *)$ in the étale topology. We have $\mathbb{Z}(0) \simeq \mathbb{Z}$ on a normal scheme and $\mathbb{Z}(1) \simeq \mathbb{G}_m[-1]$ on a smooth scheme. For a separated, noetherian scheme X of finite Krull dimension, the Mayer–Vietoris property yields the isomorphism $H^i(X, \mathbb{Z}(n)) \simeq H^i(X_{\mathrm{Zar}}, \mathbb{Z}(n))$. For X smooth, filtering $z^*(X, *)$ by codimension, we get the very useful Gersten resolution

$$0 \to \mathcal{H}^p(\mathbb{Z}(n)) \to \bigoplus_{x \in X^{(0)}} (i_x)_* H^p(k(x), \mathbb{Z}(n)) \to \bigoplus_{x \in X^{(1)}} (i_x)_* H^{p-1}(k(x), \mathbb{Z}(n-1)) \to$$

Here $X^{(s)}$ denotes the set of points in X of codimension s.

For X smooth and m invertible on X, rigidity in higher Chow groups and étale cohomology and the vanishing of $H^i(k_{\text{ét}}, \mathbb{Z}/m(n))$ for an algebraically closed field k [33] imply that $\mathbb{Z}/m(n)_{\text{ét}} \xrightarrow{\sim} \mu_m^{\otimes n}$. We get the isomorphism

$$c^{\text{ét}}_{i,n}\colon H^i(X_{\text{ét}}, \mathbb{Z}/m(n)) \xrightarrow{\sim} H^i(X_{\text{ét}}, \mu_m^{\otimes n}).$$

Conjecture 3.2 (Beilinson–Lichtenbaum). The canonical map

$$\rho_{i,n}\colon H^i(X_{\mathrm{Zar}}, \mathbb{Z}/m(n)) \to H^i(X_{\text{ét}}, \mathbb{Z}/m(n))$$

is an isomorphism for $i \leq n$.

It is clear that the Bloch–Kato conjecture is a special case of the above conjecture. What is not obvious is that it also implies it.

Theorem 3.3 (Suslin–Voevodsky [35], Geisser–Levine [16]). *The Bloch–Kato conjecture implies the Beilinson–Lichtenbaum conjecture.*

To prove this result via Gersten resolution one passes to the following statement for fields: the Bloch–Kato isomorphism $H^n(F, \mathbb{Z}/m(n)) \simeq K_n^M(F)/n \xrightarrow{\sim} H^n(F_{\text{ét}}, \mu_m^{\otimes n})$ for all fields F that are finitely generated over the base field implies that $\rho_{i,n}\colon H^i(F, \mathbb{Z}/m(n)) \to H^i(F_{\text{ét}}, \mathbb{Z}/m(n))$ is an isomorphism for all such F and $i \leq n$. This is proved by descending induction on the degree of cohomology by "bootstrapping" the Bloch–Kato isomorphism into relative cohomology of cubical complexes.

Unconditionally, we have two important results

Theorem 3.4 (Levine [24]). *If $\mu_m \in \Gamma(X, \mathcal{O}_X)$, inverting the Bott element $\beta_m \in H^0(X_{\text{Zar}}, \mathbb{Z}/m(1))$ gives an isomorphism*

$$\tilde{\rho}_{i,n}\colon H^i(X_{\text{Zar}}, \mathbb{Z}/m(n))[\beta_m^{-1}] \xrightarrow{\sim} H^i(X_{\text{ét}}, \mathbb{Z}/m(n)).$$

Theorem 3.5 (Suslin [33]). *The map $\rho_{i,n}$ is an isomorphism for X smooth over an algebraically closed field and $n \geq \dim X$.*

We can now use the Atiyah–Hirzebruch spectral sequence (3.1) and its étale analogue

$$E_2^{s,t} = H^{s-t}(X_{\text{ét}}, \mathbb{Z}/m(-t)) \Rightarrow K^{\text{ét}}_{-s-t}(X, \mathbb{Z}/m)$$

constructed by Levine to pass from motivic cohomology to K-theory and to conclude that

Theorem 3.6 (Levine [23]). *The Beilinson–Lichtenbaum conjecture implies the Quillen–Lichtenbaum conjecture.*

For X smooth over a perfect field of characteristic $p > 0$, Geisser–Levine [15] have shown that there is a quasi-isomorphism (in the Zariski and étale topology) $\mathbb{Z}/p^r(n) \xrightarrow{\sim} \nu_r^n[-n]$. They derive it from the fact that for any field k of characteristic p, $H^i(k, \mathbb{Z}/p^r(n)) = 0$ for $i \neq n$, which in turn they induce from the Bloch–Kato isomorphism $H^n(k, \mathbb{Z}/p^r(n)) \xleftarrow{\sim} K_n^M(k)/p^r \xrightarrow{\sim} \nu_r^n(k)$. As a result we get

$$H^{i+n}(X, \mathbb{Z}/p^r(n)) \simeq H^i(X_{\text{Zar}}, \nu_r^n), \quad H^{i+n}(X_{\text{ét}}, \mathbb{Z}/p^r(n)) \simeq H^i(X_{\text{ét}}, \nu_r^n).$$

The above implies that $\tilde{\pi}_n(\mathcal{K}/p^r) \simeq \nu_r^n$: via the Bloch–Lichtenbaum spectral sequence $H^{s-t}(k, \mathbb{Z}/p^r(-t)) \Rightarrow K_{-s-t}(k, \mathbb{Z}/p^r)$, the computation of $H^i(k, \mathbb{Z}/p^r(n))$ yields the isomorphism $K_n^M(k)/p^r \to K_n(k, \mathbb{Z}/p^r)$; having that it suffices now to evoke Gersten resolution.

3.2. Motivic cohomology over Dedekind domains. The construction of Bloch higher Chow groups and some of its basic properties (most notably the localization exact sequence and the Atiyah–Hirzebruch spectral sequence) as well as some computations of motivic sheaves can be extended to schemes of finite type over a Dedekind scheme ([25], [23], [14]). Here is an example. Let X be a smooth scheme over a complete discrete valuation ring V of mixed characteristic $(0, p)$ with a perfect residue field k. Denote by $i: Y \hookrightarrow X$ and $j: U \hookrightarrow X$ the special and generic fibers, respectively. We will sketch how assuming the Bloch–Kato conjecture mod p we get a quasi-isomorphism [14]

$$i^*\mathbb{Z}/p^n(r)_{\text{ét}} \to S_n(r) \quad \text{for } r < p-1.$$

Here $S_n(r)$ is the syntomic complex of Fontaine–Messing [8] (philosophically) defined as the mapping cone of the map $Ru_*J^{[r]}_{X_n/W_n(k)} \overset{1-\phi^r}{\to} Ru_*\mathcal{O}_{X_n/W_n(k)}$, where $W(k)$ is the ring of Witt vectors of k, $\mathcal{O}_{X_n/W_n(k)}$ is the crystalline structure sheaf, $J_{X_n/W_n(k)} = \ker(\mathcal{O}_{X_n/W_n(k)}) \to \mathcal{O}_{X_n})$, $u: X_n/W_n(k)_{\text{cr,ét}} \to X_{n,\text{ét}}$ is the natural projection, and $\phi^r = \phi/p^r$ is the divided Frobenius. We always get the long exact sequence

$$\to H^i(X_n, S_n(r)) \to H^i_{\text{cr}}(X_n/W_n(k), J^{[r]}) \xrightarrow{1-\phi^r} H^i_{\text{cr}}(X_n/W_n(k)) \to$$

By the theory of p-adic periods [21] we have the distinguished triangle

$$\to S_n(r) \to \tau_{\le n} i^* Rj_*\mu_{p^n}^{\otimes r} \to \nu_n^{r-1}[-r] \to$$

This triangle can be seen as a realization of the "localization" sequence for the étale motivic sheaves: we apply the above computations of motivic sheaves over fields and a purity result $\mathbb{Z}(r-1)_{\text{ét}}[-2] \xrightarrow{\sim} \tau_{\le r+1} Ri^!\mathbb{Z}(r)_{\text{ét}}$ (contingent on the Beilinson–Lichtenbaum conjecture mod p) to the localization sequence and get the distinguished triangle

$$\to i^*\mathbb{Z}/p^n(r)_{\text{ét}} \to \tau_{\le r} i^* Rj_*\mu_{p^n}^{\otimes r} \to \nu_n^{r-1}[-r] \to .$$

Comparing the above two triangles, we get that the cycle class map $i^*\mathbb{Z}/p^n(r)_{\text{ét}} \to S_n(r)$ is a quasi-isomorphism inducing a cycle map (an isomorphism for X proper)

$$c^{\text{syn}}_{i,r}: H^i(X_{\text{ét}}, \mathbb{Z}/p^n(r)) \to H^i(X, S_n(r)), \quad r < p-1.$$

4. Application: p-adic Hodge theory

In p-adic Hodge theory we attempt to understand p-adic Galois representations coming from the étale cohomology of varieties over p-adic fields via the de Rham cohomology of these varieties. The maps relating étale and de Rham cohomology groups are called p-adic period morphisms. Just as in the classical case, we would like to see them as integration of differential forms. Motivic cohomology allows us to do that [30], [31]. We will sketch briefly how.

Remark 4.1. The main comparison theorems of p-adic Hodge theory were proved earlier by two different methods: by Fontaine–Messing–Kato [8], [20], Kato [19], and Tsuji [39] via a study of p-adic nearby cycles and by Faltings [6], [7] using the theory of almost étale extensions.

4.1. The good reduction case. Let k be a perfect field of positive characteristic p, $W(k)$ the corresponding ring of Witt vectors and K its field of fractions. Let $\overline{K}$ be an algebraic closure of K and let $\mathrm{Gal}(\overline{K}/K)$ denote its Galois group. Let X be a smooth proper scheme over $V = W(k)$ of relative dimension d. We have a functor which carries the crystalline cohomology groups of X with all their structures into representations of $\mathrm{Gal}(\overline{K}/K)$. For $p-2 \geq r \geq i$, set

$$\boldsymbol{L}(H^i_{\mathrm{cr}}(X_n/V_n)\{-r\}) := (F^0(H^i_{\mathrm{cr}}(X_n/V_n)\{-r\} \otimes B^+_{\mathrm{cr},n}))^{1=\phi^0}.$$

Here $B^+_{\mathrm{cr},n} = H^*_{\mathrm{cr}}(\mathrm{Spec}(\overline{V}_n)/W_n(k))$ is one of Fontaine's rings of periods. It is equipped with a decreasing filtration $F^i B^+_{\mathrm{cr},n}$, Frobenius, and an action of the group $\mathrm{Gal}(\overline{K}/K)$. The crystalline cohomology groups $H^i_{\mathrm{cr}}(X_n/V_n) \simeq H^i_{dR}(X_n/V_n)$ have a natural Hodge filtration and ϕ^0 comes from the tensor of divided Frobeniuses $\phi^j = \phi/p^j$. The twist $\{-r\}$ refers to twisting the Hodge filtration and the Frobenius.

Conjecture 4.2 (Crystalline conjecture). For p large enough, there exists a canonical Galois equivariant period isomorphism

$$\alpha_{\mathrm{cr}} \colon H^i(X_{\overline{K}}, \mu_n^{\otimes r}) \xrightarrow{\sim} \boldsymbol{L}(H^i_{\mathrm{cr}}(X_n/V_n)\{-r\}).$$

The proof using K-theory we sketch here works for $r \geq 2d$, $p-2 \geq 2r+d$ (or rationally with no restriction on p and the degree of the finite extension $V/W(k)$).

Since $B^+_{\mathrm{cr},n} \simeq H^*_{\mathrm{cr}}(\overline{V}_n/V_n)$, by the Künneth formula $H^*_{\mathrm{cr}}(X_n/V_n) \otimes B^+_{\mathrm{cr}} \simeq H^*_{\mathrm{cr}}(X_{\overline{V},n}/V_n)$, where $\overline{V}$ is the integral closure of V in $\overline{K}$. The defining property of syntomic cohomology yields a natural map (in fact an isomorphism)

$$H^i(X_{\overline{V}}, S_n(r)) \to \boldsymbol{L}(H^i_{\mathrm{cr}}(X_n/V_n)\{-r\}).$$

It follows that to prove the conjecture, by a standard argument, it suffices to construct a Galois equivariant map

$$\alpha_{i,r} \colon H^i(X_{\overline{K}}, \mu_{p^n}^{\otimes r}) \to H^i(X_{\overline{V}}, S_n(r))$$

compatible with Poincaré duality and some cycle classes. To construct this map as

an integration we will use the following diagram

$$\begin{array}{ccc}
F^r_\gamma/F^{r+1}_\gamma K_{2r-i}(X_{\overline{V}},\mathbb{Z}/p^n) & \xrightarrow[j^*]{\sim} & F^r_\gamma/F^{r+1}_\gamma K_{2r-i}(X_{\overline{K}},\mathbb{Z}/p^n) \\
\downarrow{\scriptstyle \rho_{2r-i}} & & \downarrow{\scriptstyle \rho_{2r-i}} \\
F^r_\gamma/F^{r+1}_\gamma K^{\text{ét}}_{2r-i}(X_{\overline{V}},\mathbb{Z}/p^n) & \xrightarrow[j^*]{} & F^r_\gamma/F^{r+1}_\gamma K^{\text{ét}}_{2r-i}(X_{\overline{K}},\mathbb{Z}/p^n) \\
\downarrow{\scriptstyle c^{\text{syn}}_{r,2r-i}} & & \downarrow{\scriptstyle c^{\text{ét}}_{r,2r-i}} \\
H^i(X_{\overline{V}},S_n(r)) & \xleftarrow[\alpha_{i,r}]{} & H^i(X_{\overline{K}},\mu^{\otimes r}_{p^n}).
\end{array}$$

The right-hand side allows us to represent étale classes by higher algebraic cycle classes on $X_{\overline{K}}$. Those can be lifted (via j^*) to the integral model $X_{\overline{V}}$ and we can integrate differential forms along them to get the period map $\alpha_{i,r}$. Specifically, by Quillen–Lichtenbaum conjecture or by Suslin the map ρ_{2r-i} is an isomorphism for $2r-i \geq \operatorname{cd}_p X_{\text{ét}} = 2d$. The degeneration of the étale Atiyah–Hirzebruch spectral sequence gives that $c^{\text{ét}}_{r,2r-i}$ is an isomorphism modulo small torsion. Also the restriction j^* is an isomorphism: since the scheme $X_{\overline{V}}$ is smooth we can pass to K'-theory; by localization, the kernel and cokernel of j^* is controlled by mod p^n K'-groups of special fibers and those can be killed by totally ramified extensions of V of degree p^n. For p and r as above, we define the map $\alpha_{i,r}$ to make this diagram commute.

Corollary 4.3. *For $r \geq d + i/2$, $p-2 \geq r + d/2$, there exists a unique period map $\alpha_{i,r}\colon H^i(X_{\overline{K}},\mu^{\otimes r}_{p^n}) \to H^i(X_{\overline{V}},S_n(r))$ compatible with the étale and syntomic higher Chern classes from K-theory mod p^n of $X_{\overline{K}}$ and $X_{\overline{V}}$.*

Based on [29], [28] we expect all the existing constructions of the period maps to be compatible with higher Chern classes hence equal.

Assume that we are able to define syntomic higher cycle maps without using p-adic periods. Then a construction of the period map $\alpha_{i,r}$ as an integral can be done in a more precise way by the following diagram.

$$\begin{array}{ccc}
H^i(X_{\overline{V}},\mathbb{Z}/p^n(r)) & \xrightarrow[j^*]{\sim} & H^i(X_{\overline{K}},\mathbb{Z}/p^n(r)) \\
\downarrow{\scriptstyle \rho_{i,r}} & & \downarrow{\scriptstyle \rho_{i,r}} \\
H^i(X_{\overline{V},\text{ét}},\mathbb{Z}/p^n(r)) & \xrightarrow{j^*} & H^i(X_{\overline{K},\text{ét}},\mathbb{Z}/p^n(r)) \\
\downarrow{\scriptstyle c^{\text{syn}}_{i,r}} & & \downarrow{\scriptstyle c^{\text{ét}}_{i,r}} \\
H^i(X_{\overline{V}},S_n(r)) & \xleftarrow[\alpha_{i,r}]{} & H^i(X_{\overline{K}},\mu^{\otimes r}_{p^n})
\end{array} \tag{4.1}$$

Arguing as above, we see that the restriction map j^* is an isomorphism. The map $\rho_{i,r}$ on the right is an isomorphism for $r \geq i$ by the Beilinson–Lichtenbaum conjecture or for $r \geq d$ by Suslin. That gives the definition of $\alpha_{i,r}$ in these two cases and a proof

of the Crystalline conjecture for all i and $2d \leq r \leq p-2$. Notice that then all the maps in the above diagram are isomorphisms.

Remark 4.4. Our period map $\alpha_{i,r}$ goes in the opposite direction than the period maps constructed by other methods. This implies that one can simply use Poincaré duality to prove that the map is an isomorphism. Rationally that works well but integrally it doubles the lower bound on p.

4.2. The semistable reduction case. Let now K be a complete discrete valuation field of mixed characteristic $(0, p)$ with ring of integers V and a perfect residue field k. Let $X^\times$ be a fine and saturated log-smooth proper $V^\times$-scheme, where V is equipped with the log-structure associated to the closed point, such that the generic fiber X_K is smooth over K and the special fiber $X_0^\times$ is of Cartier type. A standard example would be a scheme X with simple semistable reduction.

Conjecture 4.5 (Semistable conjecture). There exists a natural period isomorphism

$$\alpha_{st}\colon H^*(X_{\overline{K}}, \mathbb{Q}_p) \otimes_{\mathbb{Q}_p} B_{\mathrm{st}} \simeq H^*_{\mathrm{cr}}(X_0^\times / W(k)^0) \otimes_{W(k)} B_{\mathrm{st}}$$

preservings Galois action, monodromy, filtration and Frobenius.

Here the period ring B_{st} is equipped with Galois action, Frobenius and monodromy operators. It maps naturally into another ring of periods B_{dR}, which is equipped with a decreasing filtration. The log-crystalline cohomology groups $H^*_{\mathrm{cr}}(X^\times / W(k)^0)[1/p]$ (analogues of limit Hodge structures) are also equipped with Frobenius and monodromy operators. There is also a canonical isomorphism $K \otimes_{W(k)} H^*_{\mathrm{cr}}(X_0^\times / W(k)^0) \simeq H^*_{dR}(X_K/K)$ via which Hodge filtration induces a descending filtration on these groups. The period isomorphism and its base change to B_{dR} should preserve all these structures. As a corollary, one gets that the étale cohomology as a Galois representation can be recovered from the log-crystalline cohomology

$$H^*(X_{\overline{K}}, \mathbb{Q}_p) \simeq (H^*_{\mathrm{cr}}(X_0^\times / W(k)^0) \otimes_{W(k)} B_{\mathrm{st}})^{N=0,\phi=1} \cap F^0(B_{dR} \otimes_K H^*_{dR}(X_K/K)).$$

In the above formula the kernel of the monodromy was computed by Kato to be $\mathbb{Q} \otimes \operatorname{proj\,lim}_n H^*_{\mathrm{cr}}(X^\times_{\overline{V},n} / W_n(k))$. If we now take into account both Frobenius and the filtration, we can pass to log-syntomic cohomology and we see that to prove the conjecture it suffices to construct a Galois equivariant family of maps

$$\alpha^n_{i,r}\colon H^i(X_{\overline{K}}, \mu_{p^n}^{\otimes r}) \to H^i_{\mathrm{cr}}(X^\times_{\overline{V}}, S_n(r)),$$

at least for r large enough, that is compatible with Poincaré duality and the trace map.

The main problem with trying to carry over our motivic proof of the Crystalline conjecture to this setting is that the integral model $X_{\overline{V}}$ is in general singular. It becomes then very difficult to control the kernel and cokernel of the restriction map j^*. However the singularities are rather mild (they are of toric type) and we find [32]

that every model $X^{\times}_{V'}$, for a finite extension V'/V, can be desingularized by a log-blow-up $Y^{\times} \to X^{\times}_{V'}$. Since we are blowing up only strata this desingularization does not change the log-syntomic cohomology. Obviously it does not change the étale cohomology either, so to define the maps $\alpha^n_{i,r}$ we can work with the regular models $Y^{\times}$. We have the usual "integration" diagram

$$\begin{array}{ccc} F^r_\gamma/F^{r+1}_\gamma K_{2r-i}(Y, \mathbb{Z}/p^n) & \xrightarrow{j^*} & F^r_\gamma/F^{r+1}_\gamma K_{2r-i}(Y_K, \mathbb{Z}/p^n) \\ \downarrow{\scriptstyle c^{\mathrm{syn}}_{r,2r-i}\rho_{2r-i}} & & \downarrow{\scriptstyle c^{\text{ét}}_{r,2r-i}\rho_{2r-i}} \\ H^i(Y^{\times}, S_n(r)) & \xleftarrow[\alpha^n_{i,r}]{} & H^i(Y_K, \mu_{p^n}(r)). \end{array}$$

The right-hand side of the diagram behaves like before. The restriction j^* is an isomorphism for $2r - i > \dim X_K + 1$ because by the localization sequence its kernel and cokernel are controlled by $K'_j(Y_k, \mathbb{Z}/p^n)$, which vanishes for $j > \dim X_K$ by Geisser–Levine. Hence we can integrate differential forms against higher cycles (on the integral model Y) to get the period maps $\alpha^n_{i,r}$. Again as a corollary we get a uniqueness statement for semistable period maps.

Remark 4.6. Notice that the above vanishing result of Geisser–Levine and the resulting bijectivity of the restriction map j^* are entirely p-adic phenomena. The analogous statements mod l are false. This is in contrast with the good reduction case where j^* is an isomorphism mod l as well.

Question 4.7. Is it possible to define log-motivic complexes and cohomology that would specialize to log-syntomic cohomology? More precisely, one would like to have a log-analogue of the motivic diagram (4.1) for a semistable scheme $X^{\times}$ (with logs everywhere in the left column). For that we need a good definition of log-motivic complexes $\mathbb{Z}/p^n(r)^{\times}$ and log-syntomic cycle classes

$$c^{\mathrm{syn}}_{i,r} \colon H^i(X^{\times}, \mathbb{Z}/p^n(r)^{\times}) \to H^i(X^{\times}, S_n(r))$$

(isomorphisms for X proper and $i \leq r < p - 1$). We would expect the restriction map

$$j^* \colon H^i(X^{\times}, \mathbb{Z}/p^n(r)^{\times}) \to H^i(X_K, \mathbb{Z}/p^n(r))$$

to be an isomorphism.

This question is closely related to that of the existence of limit motivic cohomology (see the recent work of Marc Levine [27] on that subject in the case of schemes over a field).

Acknowledgments. I would like to thank Thomas Geisser and Marc Levine for helpful comments.

References

[1] Bloch, S., Algebraic cycles and higher K-theory. *Adv. in Math.* **61** (3) (1986), 267–304.

[2] Bloch, S., Algebraic cycles and the Beilinson conjectures. In *The Lefschetz centennial conference*, Part I (Mexico City, 1984), Contemp. Math., 58, Amer. Math. Soc., Providence, RI, 1986, 65–79.

[3] Bloch, S., Kato, K., p-adic étale cohomology, *Inst. Hautes Études Sci. Publ. Math.* **63** (1986), 107–152.

[4] Bloch, S., Lichtenbaum, S., A spectral sequence for motivic cohomology. Preprint, 1995.

[5] Dwyer, W., Friedlander, E., Algebraic and étale K-theory. *Trans. Amer. Math. Soc.* **292** (1) (1985), 247–280.

[6] Faltings, G., Crystalline cohomology and p-adic Galois-representations. In *Algebraic analysis, geometry, and number theory* (Baltimore, MD, 1988), Johns Hopkins University Press, Baltimore, MD, 1989, 25–80.

[7] Faltings, G., Almost étale extensions. Cohomologies p-adiques et applications arithmétiques, II. *Astérisque* **279** (2002), 185–270.

[8] Fontaine, J.-M., Messing, W., p-adic periods and p-adic étale cohomology. In *Current trends in arithmetical algebraic geometry* (Arcata, Calif., 1985), Contemp. Math. 67, Amer. Math. Soc., Providence, RI, 1987, 179–207.

[9] Friedlander, E., Motivic complexes of Suslin and Voevodsky. In *Séminaire Bourbaki*, Vol. 1996/97; Astérisque **245** (1997), 5, 355–378.

[10] Friedlander, E., Grayson, D. (eds.), *Handbook of K-theory*. Springer-Verlag, New York 2005.

[11] Friedlander, E., Suslin, A., The spectral sequence relating algebraic K-theory to motivic cohomology. *Ann. Sci. École Norm. Sup.* (4) **35** (6) (2002), 773–875.

[12] Friedlander, E., Walker, M., Some remarks concerning mod-n K-theory. *Invent. Math.* **145** (3) (2001), 545–555.

[13] Fujiwara, K., A proof of the absolute purity conjecture (after Gabber). *Algebraic geometry 2000* (Azumino (Hotaka)), Adv. Stud. Pure Math. 36, Math. Soc. Japan, Tokyo 2002, 153–183.

[14] Geisser, T., Motivic cohomology over Dedekind rings. *Math. Z.* **248** (4) (2004), 773–794.

[15] Geisser, T., Levine, M., The K-theory of fields in characteristic p. *Invent. Math.* **139** (3) (2000), 459–493.

[16] Geisser, T., Levine, M., The Bloch-Kato conjecture and a theorem of Suslin-Voevodsky. *J. Reine Angew. Math.* **530** (2001), 55–103.

[17] Gillet, H., Riemann-Roch theorems for higher algebraic K-theory. *Adv. Math.* **40** (1981), 203–289.

[18] Grayson, D., Weight filtrations via commuting automorphisms. *K-Theory* **9** (2) (1995), 139–172.

[19] Kato, K., Semi-stable reduction and p-adic étale cohomology. In *Périodes p-adiques* (Bures-sur-Yvette, 1988), *Astérisque* **223** (1994), 269–293.

[20] Kato, K., Messing, W., Syntomic cohomology and p-adic étale cohomology. *Tohoku Math. J.* (2) **44** (1) (1992), 1–9.

[21] Kurihara, M., A note on p-adic étale cohomology. *Proc. Japan Acad. Ser. A Math. Sci.* **63** (7) (1987), 275–278.

[22] Levine, M., Bloch's higher Chow groups revisited. In *K-theory* (Strasbourg, 1992), Astérisque **226** (10) (1994), 235–320.

[23] Levine, M., K-theory and motivic cohomology of schemes. Preprint, 1999.

[24] Levine, M., Inverting the motivic Bott element. *K-Theory* **19** (1) (2000), 1–28.

[25] Levine, M., Techniques of localization in the theory of algebraic cycles. *J. Algebraic Geom.* **10** (2) (2001), 299–363.

[26] Levine, M., The homotopy coniveau filtration. Preprint, 2003.

[27] Levine, M., Motivic tubular neighborhoods. Preprint, 2005.

[28] Nekovar, J., Syntomic cohomology and p-adic regulators. Preprint, 1998.

[29] Nizioł, W., On the image of p-adic regulators. *Invent. Math.* **127** (2) (1997), 375–400.

[30] Nizioł, W., Crystalline conjecture via K-theory. *Ann. Sci. École Norm. Sup.* (4) **31** (5) (1998), 659–681.

[31] Nizioł, W., Semistable conjecture via K-theory. Preprint, 2003.

[32] Nizioł, W., Toric singularities: log-blow-ups and global resolutions. *J. Algebraic Geom.* **15** (2006), 1–29.

[33] Suslin, A., Higher Chow groups and étale cohomology. In *Cycles, transfers, and motivic homology theories*, Ann. of Math. Stud. 143, Princeton University Press, Princeton, NJ, 2000, 239–254.

[34] Suslin, A., On the Grayson spectral sequence. *Tr. Mat. Inst. Steklova* **241** (2003), Teor. Chisel, Algebra i Algebr. Geom., 218–253; English transl. *Proc. Steklov Inst. Math.* **241** (2003), 202–237.

[35] Suslin, A., Voevodsky, V., Bloch-Kato conjecture and motivic cohomology with finite coefficients. In *The arithmetic and geometry of algebraic cycles* (Banff, AB, 1998), NATO Sci. Ser. C Math. Phys. Sci. 548, Kluwer Acadademic Publ., Dordrecht 2000, 117–189.

[36] Thomason, R., Absolute cohomological purity. *Bull. Soc. Math. France* **112** (3) (1984), 397–406.

[37] Thomason, R., Algebraic K-theory and étale cohomology. *Ann. Sci. École Norm. Sup.* (4) **18** (3) (1985), 437–552.

[38] Thomason, R., Bott stability in algebraic K-theory. In *Applications of algebraic K-theory to algebraic geometry and number theory* (Boulder, Colo., 1983), Part I, II, Contemp. Math. 55, Amer. Math. Soc., Providence, RI, 1986, 389–406.

[39] Tsuji, T., p-adic étale cohomology and crystalline cohomology in the semi-stable reduction case. *Invent. Math.* **137** (2) (1999), 233–411.

[40] Voevodsky, V., Motivic cohomology with $Z/2$-coefficients. *Publ. Math. Inst. Hautes Études Sci.* **98** (2003), 59–104.

[41] Voevodsky, V., On motivic cohomology with $\mathbf{Z}/l$-coefficients. Preprint, 2003.

[42] Walker, M., Thomason's theorem for varieties over algebraically closed fields. *Trans. Amer. Math. Soc.* **356** (7) (2004), 2569–2648 (electronic).

Department of Mathematics, College of Science, University of Utah, Salt Lake City, Utah 84112-0090, U.S.A.

E-mail: niziol@math.utah.edu

Vanishing of L-functions and ranks of Selmer groups

Christopher Skinner and Eric Urban*

Abstract. This paper connects the vanishing at the central critical value of the L-functions of certain polarized regular motives with the positivity of the rank of the associated p-adic (Bloch–Kato) Selmer groups. For the motives studied it is shown that vanishing of the L-value implies positivity of the rank of the Selmer group. It is further shown that if the order of vanishing is positive and even then the Selmer group has rank at least two. The proofs make extensive use of families of p-adic modular forms. Additionally, the proofs assume the existence of Galois representations associated to holomorphic eigenforms on unitary groups over an imaginary quadratic field.

Mathematics Subject Classification (2000). 11F67, 11F80, 11F55, 11F33, 11F85, 11F11.

Keywords. p-adic modular forms, Galois representations, Selmer groups, L-functions.

Introduction

This paper aims to connect the order of vanishing of the L-functions of certain (motivic p-adic) Galois representations with the ranks of their associated Selmer groups. This connection, really an assertion of equality, is part of the general Bloch–Kato conjectures (cf. [BK] and [FP]), but its origins are in the 'class number formula' for number fields – part of which is the assertion that the order of vanishing at $s = 0$ of the Dedekind zeta-function $\zeta_K(s)$ of a number field K equals the rank of the group of units of K – and the celebrated conjecture of Birch and Swinnterton-Dyer – which asserts that the order of vanishing at $s = 1$ of the L-function $L(E, s)$ of an elliptic curve over a number field K equals the rank of the Mordell–Weil group $E(K)$. In both of these instances the equality can be restated in terms of ranks of Selmer groups (in the case of elliptic curves this requires finiteness of the (p-primary part of the) Tate–Shafarevich group of the curve).

In this paper we work in the context of a polarized regular (pure motivic) Galois representation $R\colon G_{\mathcal{K}} \to \mathrm{GL}_d(L)$ of the absolute Galois group $G_{\mathcal{K}}$ of an imaginary quadratic field $\mathcal{K}$ defined over a p-adic field L; we fix a prime p that splits in $\mathcal{K}$. The polarization condition is an isomorphism

$$R^\vee(1) \cong R^c$$

*Research of the first author sponsored in part by grants from the National Science Foundation and a fellowship from the David and Lucile Packard Foundation. Research of the second author sponsored by National Science Foundation grant DMS-04-01131.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of the arithmetic dual of R with the conjugate of R by the non-trivial automorphism c of $\mathcal{K}$. The (motivic and) regular condition is that R is Hodge–Tate at the primes above p and that the Hodge–Tate weights are regular. We further restrict to the case where the Hodge–Tate weights of R do not include 0 or -1. Unfortunately, this excludes the case of elliptic curves. The L-functions $L(R, s)$ of such Galois representations, defined using geometric Frobenius elements (throughout we adopt geometric conventions), are expected to have meromorphic continuations to all of $\mathbb{C}$ and to satisfy the functional equation

$$L(R, s) = \varepsilon(R, s)L(R^\vee(1), 1 - s).$$

The value $s = 0$ is a critical value of $L(R, s)$, and the connection between orders of vanishing and ranks of Selmer groups is the following.

Conjecture. $\mathrm{ord}_{s=0}L(R, s) = \mathrm{rank}_L H^1_f(\mathcal{K}, R^\vee(1))$.

Here $H^1_f(\mathcal{K}, R^\vee(1)) \subseteq H^1(\mathcal{K}, R^\vee(1))$ is the Bloch–Kato Selmer group. This is defined by imposing local conditions at all primes. At primes not dividing p the classes are required to be unramified, while at primes v dividing p they are required to be crystalline: their image in $H^1(I_v, R^\vee(1) \otimes B_{\mathrm{cris}})$ is zero, where B_{cris} is Fontaine's ring of p-adic periods.

The Galois representations we consider are expected to be automorphic in the sense that for a given R there should exist a unitary group $U(V)$ in d-variables, an automorphic representation π of $U(V)$ with infinity-type a holomorphic discrete series, and an algebraic idele class character χ of $\mathcal{K}$ satisfying $\chi|^\times_{\boldsymbol{A}_\mathbb{Q}} = |\cdot|^{2\kappa'}_{\boldsymbol{A}_\mathbb{Q}}$ such that $L(R, s) = L(\pi, \chi^{-1}, s + \kappa' + 1/2)$, where the right-hand side is a twist of the standard L-function of π. Such an identification is generally the only known strategy for proving the conjectured analytic properties of $L(R, s)$. So we start by assuming that given π and χ, the corresponding R exists. In general this is only known for unitary groups in 3 or fewer variables (see [BR92]) or under certain local hypotheses on π (see [HL04]) (these conditions certainly do not hold in all the cases we consider). We further assume that π and χ are unramified at primes above p. We then prove two theorems – Theorems 4.3.1 and Theorems 5.1.1 – in the direction of the above conjecture. We emphasize that their proofs require the existence of Galois representations associated to certain cuspidal representations of unitary groups; this existence is made precise in Conjecture 4.1.1. The first of these theorems is the following.

Theorem A. *If* $L(\pi, \chi^{-1}, \kappa' + 1/2) = L(R, 0) = 0$ *then* $\mathrm{rank}\, H^1_f(\mathcal{K}, R^\vee(1)) \geq 1$.

We include a few remarks about this theorem.

(i) In earlier work [SU02], [SU06] we proved a result similar to Theorem A: if F is a holomorphic modular form of even weight $2k$ and trivial nebentypus and ordinary for p and if $\mathrm{ord}_{s=k}L(F, s)$ is odd then the rank of the corresponding p-adic Selmer group $H^1_f(\mathbb{Q}, V_F(k))$ is positive (V_F is the p-adic Galois representation associated

to F). The positivity of the rank in the case of even order vanishing – at least if F is unramified at p – will follow from our forth-coming work [SU-MC] on the Iwasawa main conjecture for modular forms[1]. For prior results in the same vein (by Gross and Zagier, Greenberg, Nekovář, Bellaïche,…) the interested reader should consult the introduction to [SU06].

(ii) When π is just an idele class character, so R is one-dimensional, Theorem A is unconditional. Since no hypothesis is imposed on the epsilon factor $\varepsilon(R, 0)$, Theorem A in this case generalizes the complex multiplication case of [SU02], [SU06], where $\varepsilon(R, 0) = -1$ is required (the $\varepsilon(R, 0) = -1$ case is also the main result of [BC04]; our proof of Theorem A provides an alternate proof of this case).

(iii) Suppose F is a holomorphic modular form of even weight $2k > 2$, trivial nebentypus, and level prime to p. One consequence of Theorem A is that if $L(F, k) = 0$, then the rank of $H^1_f(\mathcal{K}, V_F(k))$ is positive. Choosing $\mathcal{K}$ so that the twist $F_{\mathcal{K}}$ of F by the character of $\mathcal{K}$ is such that $L(F_{\mathcal{K}}, k) \neq 0$ and appealing to a result of Kato [Ka04] that asserts $H^1_f(\mathbb{Q}, V_{F_{\mathcal{K}}}(k)) = 0$ in this case, we can then conclude that $H^1_f(\mathbb{Q}, V_F(k))$ has positive rank. This provides another proof of the results from remark (i) as well as an extension of them to the non-ordinary case.

(iv) The authors of [BC04] have announced a result in the spirit of Theorem A but with a number of additional hypotheses, including $\varepsilon(R, 0) = -1$ and certain of the Arthur conjectures.

Our proof of Theorem A follows along the same lines as the proof of the main result in [SU02], [SU06]. As explained in §1, the vanishing of the L-function at $s = 0$ implies the existence of a holomorphic Eisenstein series on a larger unitary group. This is analogous to the situation in *loc. cit.* where odd-order vanishing implies the existence of a special cuspform on a larger group, there a symplectic group of genus 2. In §§2 and 3 we construct a p-adic deformation of this Eisenstein series, a p-adic family of automorphic representations containing the Eisenstein representation. The generic member of this family is cuspidal. The Galois representations associated to these cuspidal representations (whose existence is one of our primary hypotheses) are generically irreducible. Putting all this together, we construct an irreducible family of Galois representations that specializes at one point to the reducible Galois representation $1 \oplus \varepsilon_p \oplus R$ (the Galois representation of the Eisenstein series). By a now standard argument, we then deduce the existence of a non-trivial $G_{\mathcal{K}}$-extension $0 \to L(1) \to E \to R \to 0$. Using a result of Kisin [Ki] we are able to deduce that this extension lies in $H^1_f(\mathcal{K}, R^\vee(1))$.

In the last section of this paper we extend Theorem A to a higher-rank case (under the same hypotheses on χ and π).

Theorem B. *If* $\mathrm{ord}_{s=0} L(\pi, \chi^{-1}, s + \kappa' + 1/2) = \mathrm{ord}_{s=0} L(R, s)$ *is even and positive, then* $\mathrm{rank}\, H^1(\mathcal{K}, R^\vee(1)) \geq 2$.

The proof of Theorem B relies on that of Theorem A. The hypothesis that $L(R, s)$

[1] This work includes a local hypothesis on the modular form F and its associated mod p Galois representation.

vanishes to even order at $s = 0$ means that $\varepsilon(R, 0) = 1$. And so the epsilon factor of the primitive L-function of the Eisenstein series constructed in the proof of Theorem A is equal to -1. This is then true of all the cuspidal representations in the p-adic family from the proof of that theorem. In particular, their L-functions satisfy the hypothesis of Theorem A. So running through the proof of that theorem for these cuspidal representations, one deduces the existence of a p-adic family of generically irreducible Galois representations which specializes at one point to the representation $L^2 \oplus L(1)^2 \oplus R$. And then from this we deduce the existence of a subspace of rank 2 in the Selmer group.

The proofs of Theorems A and B rely crucially on the theory of p-adic families of automorphic representations, especially as developed in [KL] and in [U06].

The authors thank Laurent Berger and Mark Kisin for some useful conversations.

Standard notation. Throughout this paper p is a fixed prime. Let $\overline{\mathbb{Q}}$ and $\overline{\mathbb{Q}}_p$ be, respectively, algebraic closures of $\mathbb{Q}$ and $\mathbb{Q}_p$ and let $\mathbb{C}$ be the field of complex numbers. We fix embeddings $\iota_\infty : \overline{\mathbb{Q}} \hookrightarrow \mathbb{C}$ and $\iota_p : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_p$. Throughout we implicitly view $\overline{\mathbb{Q}}$ as a subfield of $\mathbb{C}$ and $\overline{\mathbb{Q}}_p$ via the embeddings ι_∞ and ι_p. Let $\mathbb{C}_p$ be the completion of $\overline{\mathbb{Q}}_p$ with respect to its p-adic metric. We fix an identification $\mathbb{C}_p \cong \mathbb{C}$ compatible with the embeddings ι_p and ι_∞.

We fix $\mathcal{K} \subset \overline{\mathbb{Q}}$ an imaginary quadratic field. We denote by c the complex conjugation of $\mathbb{C}$ (and hence of $\mathcal{K}$). We assume that p splits in $\mathcal{K}$: $p = \wp\wp^c$ with $\wp$ the prime ideal of $\mathcal{K}$ induced by ι_p. We write ϖ for an uniformizer of $\wp$.

1. Eisenstein series and vanishing of L-functions

1.1. Unitary groups. Let θ be a totally imaginary element in $\mathcal{K}$ such that $-i\theta > 0$ and let $\Delta = \theta\bar{\theta}$ (a positive rational number). In Sections 2–5 we will assume that $\mathrm{ord}_p(\Delta) = 0$. Given integers $b \geq a \geq 0$, $a + b = d > 0$, we let

$$T_{a,b} = \begin{pmatrix} & & 1_b \\ & \theta^{-1} & \\ -1_b & & \end{pmatrix} \in \mathrm{GL}_d(\mathcal{K}).$$

We let $G_{a,b}$ be the unitary group associated to this (skew-Hermitian) matrix: for any $\mathbb{Q}$-algebra R

$$G_{a,b}(R) = \{g \in \mathrm{GL}_d(\mathcal{K} \otimes R) \ : g T_{a,b} {}^t\bar{g} = T_{a,b}\}.$$

Then $G_{a,b}(\mathbb{R})$ is a real unitary group of signature (a, b). The unbounded symmetric domain associated to this group is

$$\mathcal{D}_{a,b} = \Big\{ \begin{bmatrix} z \\ u \\ 1_a \end{bmatrix} \in \mathrm{M}_{d\times a}(\mathbb{C}) \ : \ z \in \mathrm{M}_{a\times a}(\mathbb{C}),\ u \in \mathrm{M}_{(b-a)\times a}(\mathbb{C}), \\ \theta^{-1}(z - z^*) - u^*u > 0 \Big\}.$$

The action of $G_{a,b}(\mathbb{R})$ on $\mathcal{D}_{a,b}$ is defined as follows: for $g \in G_{a,b}(\mathbb{R})$ and $x \in \mathcal{D}_{a,b}$

$$g(x) = g \cdot x \cdot t^{-1}, \quad g.x = \begin{bmatrix} r \\ s \\ t \end{bmatrix}, \ r, t \in \mathrm{M}_{a\times a}(\mathbb{C}),$$

where . denotes the usual matrix multiplication. Let

$$x_0 = \begin{bmatrix} i \\ 0 \\ 1 \end{bmatrix} \in \mathcal{D}_{a,b}.$$

The stabilizer of x_0 in $G_{a,b}(\mathbb{R})$ is a maximal compact, which we denote $K_{a,b}$. This is the group of $\mathbb{R}$-points of an $\mathbb{R}$-group that we also denote by $K_{a,b}$. The map $g \mapsto g(x_0)$ is a real analytic isomorphism of $G_{a,b}(\mathbb{R})/K_{a,b}$ with $\mathcal{D}_{a,b}$. We will often write an element g of $G_{a,b}$ or $\mathrm{M}_{d\times d}$ in block form: $g = (g_{ij})_{1\le i,j\ge 3}$ with $g_{11}, g_{33} \in \mathrm{M}_{a\times a}$. We let $B_{a,b}$ be the $\mathbb{Q}$-rational Borel of $G_{a,b}$ defined by requiring $g_{21} = g_{31} = g_{32} = 0$ and g_{33} to be upper-triangular (so g_{11} is lower-triangular).

Let

$$c = c_{a,b} = 2^{-1/2} \begin{pmatrix} 1_a & & -i1_a \\ & \sqrt{|\theta|}1_{b-a} & \\ -i1_a & & 1_a \end{pmatrix} \in \mathrm{GL}_d(\mathbb{C}).$$

Then $cT_{a,b}{}^t\bar{c} = i/2\mathrm{diag}(1_a, -1_b)$, so $k \mapsto ckc^{-1}$ identifies $K_{a,b}$ with the $\mathbb{R}$-group $U(a) \times U(b)$ (embedded diagonally in $\mathrm{GL}_d(\mathbb{C})$). Let $H_{a,b}$ be the Cartan subgroup of $K_{a,b}$ that is identified with the group of diagonal matrices in $U(a) \times U(b)$. Let $J_{a,b}\colon G_{a,b}(\mathbb{R}) \times \mathcal{D}_{a,b} \to K_{a,b}(\mathbb{C})$ be the canonical automorphy factor: if $k \in K_{a,b}$ then $J_{a,b}(k, x_0) = k$, and

$$cJ_{a,b}\Big(\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ & a_{22} & a_{23} \\ & & a_{33} \end{pmatrix}, x_0\Big)c^{-1} = \big(a_{11}, \begin{pmatrix} a_{22} & a_{23} \\ & a_{33} \end{pmatrix}\big).$$

These properties, together with the usual cocycle condition, completely determine $J_{a,b}$.

We also fix for each prime ℓ a maximal compact $K_{a,b,\ell} \subset G_{a,b}(\mathbb{Q}_\ell)$, and let $K_{a,b,f} = \prod K_{a,b,\ell}$.

Let $a' = a+1$, $b' = b+1$. Let $P_{a,b}$ be stabilizer in $G_{a',b'}$ of the line $\{(0, \dots, 0, x) \in \mathcal{K}^{d+2} : x \in \mathcal{K}\}$. Then $P_{a,b}$ is a standard, maximal $\mathbb{Q}$-parabolic of $G_{a',b'}$ with standard Levi subgroup $L_{a,b}$ isomorphic to $G_{a,b} \times \mathrm{Res}_{\mathcal{K}/\mathbb{Q}}\boldsymbol{G}_m$: a pair $(g, t) \in G_{a,b} \times \mathrm{Res}_{\mathcal{K}/\mathbb{Q}}\boldsymbol{G}_m$ is identified with

$$m(g,t) = \begin{pmatrix} g_{11} & & g_{12} & g_{13} \\ & \bar{t}^{-1} & & \\ g_{21} & & g_{22} & g_{23} \\ g_{31} & & g_{32} & g_{33} \end{pmatrix}_{t} \in G_{a',b'}.$$

We write $N_{a,b}$ for the unipotent radical of $P_{a,b}$. The map $r_{a,b} : \mathcal{D}_{a',b'} \to \mathcal{D}_{a,b}$ given by

$$r_{a,b}\Big(\begin{bmatrix} z \\ u \\ 1_{a'} \end{bmatrix}\Big) = \begin{bmatrix} z' \\ u' \\ 1_a \end{bmatrix},$$

$$z = (z_{ij})_{1\le i,j\le a+1}, \quad z' = (z_{ij})_{1\le i,j\le a},$$

$$u = (u_{i,j}), \quad u' = (u_{i,j})_{j\le a},$$

is $P_{a,b}(\mathbb{R})$-equivariant in the sense that if $p = m(g,t)n \in P_{a,b}(\mathbb{R}) = L_{a,b}(\mathbb{R})N_{a,b}(\mathbb{R})$ then $r_{a,b}(p(x)) = g(r_{a,b}(x))$.

The algebraic characters of $G_{a,b}$ correspond to d-tuples of integers $(c_d, \dots, c_{b+1}; c_1, \dots, c_b)$ in the usual way. The irreducible algebraic representations of $K_{a,b}$ are then classified by those d-tuples satisfying $c_1 \geq c_2 \geq \cdots \geq c_b$ and $c_{b+1} \geq c_{b+2} \geq c_d$. Such d-tuples also classify the L-packets of discrete series representations of $G_{a,b}(\mathbb{R})$. The holomorphic discrete series correspond to those d-tuples such that $c_b - c_{b+1} \geq d$. Given such a d-tuple τ, we write π_τ^H for the corresponding holomorphic discrete series.

When a and b are fixed or their exact values unimportant, we write G for $G_{a,b}$ and H for $G_{a',b'}$. In our remaining notation we drop the subscript 'a, b' and replace the subscript 'a', b'' with a superscript $'$.

1.2. *L*-functions. Let π be an automorphic representation of G and χ an idele class character of $\boldsymbol{A}_{\mathcal{K}}^\times$. We write $L(\pi, \chi, s)$ for the standard L-function associated to π and χ: if $\mathrm{BC}(\pi)$ is the formal base change of π to $\mathrm{GL}(d)_{/\mathcal{K}}$ then $L(\pi, \chi, s) = L(\mathrm{BC}(\pi), \chi, s)$. If S is a finite set of places of $\mathbb{Q}$ then the superscript 'S' on $L^S(\pi, \chi, s)$ will, as usual, mean that the Euler factors at the places in S have been omitted. If $d > 1$ and π is cuspidal and not endoscopic or CAP, then $\mathrm{BC}(\pi)$ is expected to be cuspidal, hence the L-functions $L^S(\pi, \chi, s)$ are expected to satisfy the following:

$$L^S(\pi, \chi, s) \text{ is holomorphic on all of } \mathbb{C} \tag{1.2.1}$$

and

$$\text{if } \pi \text{ and } \chi \text{ are unitary, then } L^S(\pi, \chi, s) \neq 0 \text{ for } \mathrm{Re}(s) \geq 1. \tag{1.2.2}$$

Remark 1.2.1. That $\mathrm{BC}(\pi)$ is cuspidal as expected is known in certain cases: (1) if $d = 2, 3$ or (2) if π_v is supercuspidal for some finite place v.

1.3. Eisenstein series. Given a cuspidal automorphic representation π of G with underlying space V_π and an idele class character χ of $\boldsymbol{A}_{\mathcal{K}}^\times$, we let $\rho = \rho_{\pi,\chi}$ be the representation of $P(\boldsymbol{A})$ on V_π defined by $\rho(m(g,t)n)v = \chi(t)\pi(g)v$, $m(g,t) \in M(\boldsymbol{A})$, $n \in N(\boldsymbol{A})$. Let $I(\rho)$ be the space of smooth, K'-finite functions $f : H(\boldsymbol{A}) \to V_\pi^{\mathrm{sm}}$ such that $f(pg) = \rho(p)f(g)$. We assume that V_π has been identified with a cuspidal subspace of $L^2(G(\mathbb{Q})\backslash G(\boldsymbol{A}))$, so the smooth vectors V_π^{sm} are smooth functions and the smooth, K-finite vectors $V_\pi^{\mathrm{sm,fin}}$ are cuspforms. Then evaluation at the identity converts $f \in I(\rho)$ into a $\mathbb{C}$-valued function on $H(\boldsymbol{A})$; we often write $f(x)$ for $f(x)(1)$. Bearing this in mind, given $f \in I(\rho)$ and a complex number s we consider the Eisenstein series

$$E(f; s, g) = \sum_{\gamma \in P(\mathbb{Q})\backslash H(\mathbb{Q})} f(\gamma g)\delta(\gamma g)^{s+1/2},$$

where δ is the usual modulus function for P: $\delta(m(g,t)) = |t\bar{t}|_{\boldsymbol{A}}^{-(d+1)}$. If $\mathrm{Re}(s)$ is sufficiently large (if π and χ are unitary and π is tempered then $\mathrm{Re}(s) > 1/2$ suffices)

then this series converges absolutely and uniformly for s and g in compact sets and so is holomorphic in s and defines an automorphic form on $H(\boldsymbol{A})$. The general theory of Eisenstein series provides a meromorphic continuation of $E(f; s, g)$ to all of $\mathbb{C}$.

1.4. Holomorphy and vanishing of L-functions. Suppose that $\pi = \otimes\pi_v$ is such that $\pi_\infty = \pi_\tau^H$ for some d-tuple $\tau = (c_d, \ldots, c_{b+1}; c_1, \ldots, c_b)$. We identify τ with the corresponding algebraic representation of K and write V_τ for the complex points of the underlying module (so V_τ is a finite-dimensional complex vector space and τ defines an action of $K(\mathbb{C})$ on V_τ). Then $(V_{\pi_\infty}^{\mathrm{sm,fin}} \otimes V_\tau)^K$ is one-dimensional. Let φ_∞ be a non-zero generator of this space. Let $\varphi_f \in \otimes_{\ell\neq\infty} V_{\pi_\ell}^{\mathrm{sm}}$ and let $\varphi = \varphi_\infty \otimes \varphi_f \in V_\pi^{\mathrm{sm,fin}} \otimes V_\tau$. We convert φ into something more classical as follows. We write $\tau(g, x)$ for $\tau(J(g, x))$ and set

$$F(Z) = \tau(g, x_0)\varphi(gx), \quad g \in G(\mathbb{R}),\ g(x_0) = Z \in \mathcal{D}_G.$$

This is a holomorphic function of Z, and if $U \subseteq G(\boldsymbol{A}_f)$ is an open compact such that $\varphi(gk) = \varphi(g)$ for all $k \in U$ then F satisfies

$$F(\gamma(Z)) = \tau(\gamma, Z)F(Z), \quad \gamma \in \Gamma = G(\mathbb{Q}) \cap U.$$

Let $\chi = \otimes\chi_v$ be an idele class character of $\boldsymbol{A}_{\mathcal{K}}^\times$ such that $\chi_\infty = z^n\bar{z}^m$ with $n + m$ having the same parity as d. Let $\kappa, \kappa' \in \frac{1}{2}\mathbb{Z}$ be defined by $2\kappa = n - m$ and $2\kappa' = n + m$ and assume κ satisfies

$$c_b \geq \kappa + d/2 + 1, \quad \kappa - d/2 - 1 \geq c_{b+1}. \tag{1.4.1}$$

Let ξ be the $d+2$-tuple $\xi = (c_d, \ldots, c_{b+1}, \kappa - d/2 - 1; c_1, \ldots, c_b, \kappa + d/2 + 1)$. As in the case of τ, we identify ξ with the corresponding algebraic representation of K' and write V_ξ for the complex points of the underlying module. The representation τ appears with multiplicity one in the restriction of ξ to K, the latter viewed as a subgroup of K' via $k \mapsto m(k, 1)$; the other irreducible representations appearing in this restriction have highest weight dominated by τ. We fix a K-equivariant inclusion of V_τ into V_ξ (explicitly, if v and w are respective highest weight vectors of these representations then $v \mapsto w$ determines such an inclusion). Then $(V_\pi^{\mathrm{sm,fin}} \otimes V_\xi)^K = (V_\pi^{\mathrm{sm,fin}} \otimes V_\tau)^K$ since τ is the minimal K-type in π_∞.

There are compatible factorizations $\rho = \otimes\rho_v$ and $I(\rho) = \otimes I(\rho_v)$, with $\rho_v = \rho_{\pi_v, \chi_v}$ and $I(\rho_v)$ defined similarly to ρ and $I(\rho)$. Let $\rho_f = \otimes_{v\neq\infty}\rho_v$ and $I(\rho_f) = \otimes_{v\neq\infty} I(\rho_v)$. A straight-forward application of Frobenius reciprocity shows that $(I(\rho_\infty) \otimes V_\xi)^{K'}$ is one-dimensional. Let Φ_∞ be a generator of this space. Let $\Phi_f \in I(\rho_f)$ and let $\Phi = \Phi_\infty \otimes \Phi_f \in (I(\rho) \otimes V_\xi)^{K_{H,\infty}}$. For $h \in H(\boldsymbol{A}_f)$, $\Phi(h) \in (V_\pi^{\mathrm{sm,fin}} \otimes V_\xi)^{K_{G,\infty}} = (V_\pi^{\mathrm{sm,fin}} \otimes V_\tau)^{K_{G,\infty}}$. Let $\varphi_h = \Phi(h)$. Then $\varphi_h = \varphi_\infty \otimes \varphi_{h,f}$.

We relate Φ to something more classical as we did φ (and hence each φ_h). For $g \in H(\mathbb{R})$ and $Z \in \mathcal{D}'$ we let $\xi(g, Z) = \xi(J'(g, Z))$. For $h \in H(\boldsymbol{A}_f)$ and $s \in \mathbb{C}$ we

then set

$$\mathcal{F}_h(s, Z) = \xi(g, x_0)\Phi(gh)\delta(g)^{s+1/2}, \quad g \in H(\mathbb{R}), g(x_0) = Z \in \mathcal{D}'.$$

If $U \subseteq H(\boldsymbol{A}_f)$ is an open compact such that $\Phi_f(gkh) = \Phi_f(gh)$ for all $k \in U$, then $\mathcal{F}_h$ satisfies

$$\mathcal{F}_h(s, p(Z)) = \xi(p, Z)\mathcal{F}_h(s, Z), \quad p \in P(\mathbb{Q}) \cap U.$$

It follows from the definition of $\mathcal{F}_h(s, Z)$ that if $p = m(g, t)n \in P(\mathbb{R})$ is such that $p(x_0) = Z$ (such a p always exists since $H(\mathbb{R}) = P(\mathbb{R})K_H$), then

$$\mathcal{F}_h(s, Z) = (t\bar{t})^{1/2+\kappa'-s(d+1)}F_h(r(Z)),$$

where F_h is the function on $\mathcal{D}$ associated to $\varphi_h = \Phi(h) \in (V_\pi^{\mathrm{sm,fin}} \otimes V_\xi)^K$ as above. In particular, if

$$s_0 = (1/2 + \kappa')/(d + 1)$$

then $\mathcal{F}_h(s_0, Z)$ is visibly holomorphic as a function on $\mathcal{D}'$.

Let $v_1, \ldots, v_n$ be a basis for V_ξ and write $\Phi_\infty = \sum \Phi_{\infty,i} \otimes v_i$ with $\Phi_{\infty,i} \in I(\rho_\infty)$. Put $\Phi_i = \Phi_{\infty,i} \otimes \Phi_f$ and $E(\Phi; s, g) = \sum E(\Phi_i; s, g) \otimes v_i$ and

$$E(\mathcal{F}_h; s, Z) = \xi(g, x_0)E(\Phi; s, gh), \quad g \in H(\mathbb{R}),\ g(x_0) = Z.$$

This last is a V_ξ-valued Eisenstein series on $\mathcal{D}'$. It is holomorphic at $s \in \mathbb{C}$ if $E(\Phi; s, g)$ (so if each $E(\Phi_i; s, g)$ is).

Proposition 1.4.1. *Suppose π is the twist of a tempered representation and suppose* (1.2.1) *and* (1.2.2) *hold.*

(i) *The series $E(\mathcal{F}_h; s, Z)$ is holomorphic as a function of s at $s = s_0$.*

(ii) *If $\chi|_{\boldsymbol{A}_\mathbb{Q}^\times} \neq |\cdot|_{\boldsymbol{A}}^{2\kappa'}$ or if $L(\pi, \chi^{-1}, 1/2+\kappa') = 0$ then $E(\mathcal{F}_h; Z) = E(\mathcal{F}_h; s_0, Z)$ is holomorphic as a function of Z.*

Remark 1.4.2. An important observation is that no hypotheses have been imposed on the section Φ_f. In practice we will assume π and χ to be unramified at p and take the p-component of Φ_f to be a certain 'p-stabilization' of the spherical vector, chosen to be amenable to methods of p-adic deformations of modular forms. At the primes different from p we will generally take Φ_f to be as unramified as possible.

Proof. We briefly indicate a proof of Proposition 1.4.1. Parts (i) and (ii) both follow from analyzing the constant terms of the Eisenstein series $E(\Phi_i; s, g)$ and $E(\Phi; s, g)$. First we note that since π is cuspidal and P is maximal, the constant term along a standard parabolic other than P or G is zero. The constant term of $E(\Phi_i; s, g)$ along P can be expressed in terms of the image of Φ_i under a certain intertwining operator, as we now recall.

Implicit in the factorization $V_\pi = \otimes V_{\pi_v}$ is the choice of a new vector $\phi_v \in V_{\pi_v}^{K_v}$ at each v at which π_v is unramified. If π_v and χ_v are both unramified then we let $\Phi_v^{\mathrm{sph}} \in I(\rho_v)^{K_v'}$ be the generator such that $\Phi_v^{\mathrm{sph}}(1) = \phi_v$. The factorization $I(\rho) = \otimes I(\rho_v)$ is with respect to the Φ_v^{sph}'s.

Let $\rho^\vee$ and $I(\rho^\vee)$ be defined as ρ and $I(\rho)$ were but with χ replaced by $\chi^\vee = (\chi^c)^{-1}$, and let $\rho_v^\vee$ and $I(\rho_v^\vee)$, v a place of $\mathbb{Q}$, be similarly defined. If π_v and χ_v (and hence $\chi_v^\vee$) are unramified at v, then we let $\Phi_v^{\vee,\mathrm{sph}} \in I(\rho_v^\vee)^{K_v'}$ be such that $\Phi_v^{\vee,\mathrm{sph}}(1) = \phi_v$. We let $\Phi_\infty^\vee \in (I(\rho_\infty^\vee) \otimes V_\xi)^{K'}$ be a non-zero generator and write $\Phi_\infty^\vee = \sum \Phi_{\infty,i}^\vee \otimes v_i$.

For $\phi \in I(\rho)$ or $I(\rho^\vee)$ and $s \in \mathbb{C}$ we let $\phi_s = \phi\delta^{s+1/2}$. The constant term of $E(\Phi_i; s, g)$ along P is $\Phi_{i,s} + M(s, \Phi_i)_{-s}$ where $M(s, -)\colon I(\rho) \to I(\rho^\vee)$ is the usual intertwining operator associated to P; this is meromorphic as a function of s and for $\mathrm{Re}(s)$ sufficiently large it is defined by the integral

$$M(s, \varphi)_{-s}(g) = \int_{N(\boldsymbol{A})} \varphi_s(wng)dn, \quad w = \begin{pmatrix} 1_a & & \\ & & 1 \\ & 1_b & \\ -1 & & \end{pmatrix} \in H(\mathbb{Q}). \tag{1.4.1}$$

We let $M_v(s, -)\colon I(\rho_v) \to I(\rho_v^\vee)$ be the usual local intertwining operator associated to P; for $\mathrm{Re}(s)$ sufficiently large, but independent of v, these are given by the local versions of the integral (1.4.1). If $\varphi = \otimes\varphi_v$ we then have $M(s, \varphi) = \otimes M_v(s, \varphi_v)$, provided the right-hand side converges. For us, the important properties of the $M_v(s, -)$'s are

(a) if π_v and χ_v are unramified then

$$M_v(s, \Phi_v^{\mathrm{sph}}) = \frac{L(\pi_v, \chi_v^{-1}, (d+1)s)L(\chi_v', (2d+2)s)}{L(\pi_v, \chi_v^{-1}, (d+1)s+1)L(\chi_v', (2d+2)s+1)}\Phi_v^{\vee,\mathrm{sph}},$$

where $\chi_v' = \chi_v^{-1}|_{\mathbb{Q}_v^\times}$;

(b) for a finite place v, $M_v(s, -)$ is holomorphic at $s = s_0$;

(c) $\sum M_\infty(s, \Phi_{\infty,i}) \otimes v_i = c(s)\Phi_\infty^\vee$, where $c(s)$ is a meromorphic function with a simple zero at $s = s_0$.

Part (a), of course, is a well-known computation. Part (b) follows from [Sh] and the hypothesis that π is a twist of a tempered representation. Part (c) is a relatively straight-forward computation.

Suppose $\Phi_f = \otimes\Phi_\ell$; we may assume this without loss of generality since any Φ_f is a linear combination of such. Let S be the set of primes ℓ such that π_ℓ or χ_ℓ is ramified or $\Phi_\ell \neq \Phi_\ell^{\mathrm{sph}}$. From (a) and (c) above it follows that for $\mathrm{Re}(s)$ sufficiently large we then have

$$M(s, \Phi_i) = \frac{c(s)L^S(\pi, \chi^{-1}, (d+1)s)L^S(\chi', (2d+2)s)}{L^S(\pi, \chi^{-1}, (d+1)s+1)L^S(\chi', (2d+2)s+1)} \cdot \Phi_{\infty,i}^\vee \otimes_{\ell\notin S} \Phi_\ell^{\mathrm{sph}} \otimes_{\ell\in S} M_\ell(s, \Phi_\ell).$$

Note that $\chi' = \chi^{-1}|_{\boldsymbol{A}_{\mathbb{Q}}^\times}$ is an idele class character of $\boldsymbol{A}_{\mathbb{Q}}^\times$ with infinity type $z^{-2\kappa'}$. Thus $L^S(\chi', (2d+2)s)$ is holomorphic at $s = s_0$ unless $\chi' = |\cdot|_{\boldsymbol{A}}^{-2\kappa'}$ in which case the L-function has a simple pole at $s = s_0$. It also follows that $L^S(\chi', (2d+2)s+1)$ is holomorphic and non-zero at $s = s_0$. In particular, $c(s)L^S(\chi', (2d+2)s)/L^S(\chi', (2d+2)s+1)$ is holomorphic at $s = s_0$ and non-zero only if $\chi' = |\cdot|_{\boldsymbol{A}}^{-2\kappa'}$. It follows from (1.2.1) and (1.2.2) that $L^S(\pi, \chi^{-1}, (d+1)s)/L^S(\pi, \chi^{-1}, (d+1)s+1)$ is holomorphic at $s = s_0$ and zero if and only if $L^S(\pi, \chi^{-1}, 1/2+\kappa') = 0$. Putting all this together with (b) above we find that

(d) $M(s, \Phi_i)$ is holomorphic at $s = s_0$;

(e) $M(s, \Phi_i) = 0$ if $\chi' \neq |\cdot|_{\boldsymbol{A}}^{-2\kappa'}$ or if $L(\pi, \chi^{-1}, 1/2+\kappa') = 0$.

The general theory of Eisenstein series implies that $E(\Phi_i; s, g)$ is holomorphic at $s = s_0$ if its constant terms are. Thus (d) above implies the holomorphy of $E(\Phi_i; s, g)$, and hence of each $E(\mathcal{F}_h; s, Z)$, at $s = s_0$, proving part (i) of the proposition. It also follows from the general theory of Eisenstein series that $E(\mathcal{F}_h; s_0, Z)$ is holomorphic as a function of Z if its constant terms are. This is equivalent to the holomorphy of the functions

$$Z \mapsto \xi(g, x_0)\left(\Phi_s(gx) + M(s, \Phi)_{-s}(gx)\right), \quad g \in H(\mathbb{R}),\ g(x_0) = Z,\ x \in H(\boldsymbol{A}_f),$$

where $M(s, \Phi) = \sum M(s, \Phi_i) \otimes v_i$. If $\chi' \neq |\cdot|_{\boldsymbol{A}}^{-2\kappa'}$ or $L(\pi, \chi^{-1}, 1/2+\kappa') = 0$, it follows from (e) above that this function equals $\mathcal{F}_x(s_0, Z)$ at $s = s_0$ and so is holomorphic. This proves part (ii) of the proposition. □

2. *p*-adic deformations of automorphic representations

It is impossible to list here all the contributors to this area. However, we want to emphasize that the important recent developments grew from the seminal ideas of Hida, Coleman, Mazur and Stevens. For our application, we rely mostly on an approach that has been stressed in [U06]: instead of constructing a space interpolating spaces of automorphic forms, one directly studies the p-adic properties of the 'trace' distribution. This approach is analogous to Wiles' introduction of pseudo-representations for the study of deformations of Galois representations.

2.1. Hecke operators. In this paper we take a Hecke operator to be a compactly supported smooth $\mathbb{Q}$-valued function on $G(\boldsymbol{A}_f)$. We fix a Haar measure on $G(\boldsymbol{A}_f)$ such that the maximal compact K_f has volume 1. If (π, V_π) is a smooth representation, then the action of a Hecke operator on V_π is defined using this Haar measure.

We need to restrict attention to Hecke operators of specific types at the prime p. To describe these we first fix an isomorphism $G(\mathbb{Q}_p) \cong \mathrm{GL}_d(\mathcal{K}_\wp) = \mathrm{GL}_d(\mathbb{Q}_p)$ so that $g = (g_{ij}) \in G(\mathbb{Q}_p)$ is identified with $g' = (g'_{ij}) \in \mathrm{GL}_d(\mathcal{K}_\wp)$ with $g'_{11} = {}^t g_{11}$ and

$g'_{33} = g_{33}$ and so that $B(\mathbb{Q}_p)$ is identified with a standard parabolic of $\mathrm{GL}_d(\mathbb{Q}_p)$ (i.e., contains the subgroup of upper-triangular matrices). We assume that the maximal compact $K_p \subset G(\mathbb{Q}_p)$ is identified with $\mathrm{GL}_d(\mathbb{Z}_p)$.

For each positive integer m we let $I_m \subset \mathrm{GL}_d(\mathbb{Z}_p)$ be the subgroup of matrices that are upper-triangular modulo p^m. Let $t = (t_1, \ldots, t_d)$ be a decreasing sequence of n integers. We denote by u_t the characteristic function on $\mathrm{GL}_d(\mathbb{Q}_p)$ of the double class $I_m.\mathrm{diag}(p^{t_1}, \ldots, p^{t_n}).I_m$. The u_t's generate a commutative algebra[2] via the convolution product. We denote this algebra by $\mathcal{U}_p$.

Let S be a set of finite primes containing p and the primes at which G is ramified. We let $K^S = \prod_{\ell \notin S} K_\ell \subset G(\boldsymbol{A}_f^S)$, a maximal compact open subgroup, and we put

$$\mathcal{R}_{S,p} := \mathcal{C}_c^\infty(K^S \backslash G(\boldsymbol{A}_f^S)/K^S), \mathbb{Z}) \otimes \mathcal{U}_p.$$

This Hecke operator acts naturally on any $V_\pi^{I_n.K^S}$.

2.2. *p*-stabilizations and normalizations. Let (π, V_π) be an automorphic representation such that $V_\pi^{K^S.I_n} \neq 0$ and $\pi_\infty \cong \pi_\tau^H$ with τ a d-tuple as in §1.1. There is a natural action of $R_{S,p}$ on the subspace $V_\pi^{K^S.I_n}$. The choice of an eigenspace is called a p-stabilization of π. Given an eigenspace, we write λ_π for the corresponding character of $R_{S,p}$. Of course, the choice of a p-stabilization is purely local at p: it depends only on the choice of an eigenvector for $\mathcal{U}_p$ in $\pi_p^{I_m}$.

For any $\tau = (c_d, \ldots, c_{b+1}; c_1, \ldots, c_b)$ as in §1.1, the associated normalized weight is $w_\tau := (c_1 - a, \ldots, c_b - a, c_{b+1} + b, \ldots, c_d + b)$; this defines a dominant weight of the diagonal torus of $\mathrm{GL}_d(\mathbb{Q}_p)$ since $c_b - c_{b+1} \geq d$. For the purpose of p-adic variation we normalize the character λ_π, setting $\lambda_\pi^\dagger(f) = \lambda_\pi(f)$ for any $f \in \mathcal{C}_c^\infty(K_m^S \backslash G(\boldsymbol{A}_f^S)/K_m^S), \mathbb{Z})$ and

$$\lambda_\pi^\dagger(u_t) := \frac{\lambda_\pi(u_t)}{w_\tau(t)}$$

for any $u_t \in \mathcal{U}_p$. It can be checked (cf. [Hi04]) that this normalization preserves the p-integrality of the eigenvalues.

Given a p-stabilization of π, we let I_π be the distribution defined by

$$\mathcal{C}_c^\infty(G(\boldsymbol{A}_f^p), \mathbb{Z}) \otimes \mathcal{U}_p \ni f \otimes u_t \mapsto I_\pi(f) := \mathrm{tr}(\pi(f))\lambda_\pi(u_t),$$

and we define the normalized distribution $I_\pi^\dagger$ by replacing λ_π with $\lambda_\pi^\dagger$. We call $I_\pi^\dagger$ a p-stabilized distribution associated to π. Note that $I_\pi^\dagger|_{\mathcal{R}_{S,p}} = \lambda_\pi^\dagger$.

Let $T_d \subset \mathrm{GL}_d$ be the diagonal torus and $B_d \subset \mathrm{GL}_d$ the Borel subgroup of upper-triangular matrices. Assume that $\pi_p = I(\chi) := \mathrm{Ind}_{B_d(\mathbb{Q}_p)}^{\mathrm{GL}_d(\mathbb{Q}_p)} \chi$ with χ an unramified character of $T_d(\mathbb{Q}_p)$. Let $I = I_1$. The choice of a p-stabilization is given by the choice

[2] It is easily checked that this algebra is independent of m.

of an eigenvector for $\mathcal{U}_p$ in $I(\chi)^I$. For each element of the Weyl group $W(G,T)$, there exists such an eigenvector $v_{\chi,w} \in I(\chi)^I$ with the property that

$$u_t.v_{\chi,w} = \chi^w\rho(t)\cdot v_{\chi,w},$$

where $\rho = \left(\frac{d-1}{2}, \frac{d-3}{2}, \ldots, \frac{1-d}{2}\right)$ is half the sum of the positive roots. The choice of a p-stabilization is therefore equivalent to an ordering of the Langlands parameters of the spherical representation $I(\chi)$. If $(\alpha_1, \ldots, \alpha_d)$ is the corresponding ordering, then we have

$$\lambda_\pi(u_t) = \prod_{i=1}^{d} \alpha_i^{t_i}.$$

In general, if π_p is spherical but associated to a non-unitary character χ, it may not equal the full induction of χ, in which case some orderings of the Langlands parameters do not have a corresponding p-stabilizations (see [SU02] for an example in the symplectic case).

A p-stabilization has *finite slope* if there is a d-tuple $s(\lambda_\pi^\dagger) = (s_1, \ldots, s_n) \in \mathbb{Q}^n$ such that

$$v_p(\lambda_\pi^\dagger(u_t)) = -\sum_{k=1}^{d} t_k.s_k, \quad t = \mathrm{diag}(t_1, \ldots, t_d).$$

Such a d-tuple is necessarily unique and called the slope of the p-stabilization. The integrality of the normalization implies that $s(\lambda_\pi^\dagger)$ belongs to the positive obtuse cone (in more automorphic terms this means that the Newton polygon lies above the Hodge polygon), and it can be easily checked that $s_1 + \cdots + s_d = 0$ (i.e., the Newton and Hodge polygon meet at their beginning and end) by considering the action of the center. If $s(\lambda_\pi^\dagger) = (0, \ldots, 0)$ then the p-stabilization is said to be ordinary. In general, the slope is said to be non-critical if $s_{i+1} - s_i < c_{i+1} - c_i + 2$ for all $i = 1, \ldots, d-1$. Otherwise, it is said to be critical. Note that the non-critical conditions define an alcove of the obtuse cone.

2.3. Families. We consider $\mathfrak{X}_{/\mathbb{Q}_p}$, the rigid analytic variety over $\mathbb{Q}_p$ such that

$$\mathfrak{X}(L) = \mathrm{Hom}_{\mathrm{cont}}(T(\mathbb{Z}_p), L^\times)$$

for any finite extension L of $\mathbb{Q}_p$. A point (or p-adic weight) $w \in \mathfrak{X}(L)$ is called arithmetic if the restriction of w to some open subgroup of $T(\mathbb{Z}_p)$ is algebraic and dominant. The corresponding algebraic character is then denoted $w^{\mathrm{alg}} = (a_1, \ldots, a_d)$ and we write $\mathfrak{X}(\overline{\mathbb{Q}}_p)^{\mathrm{alg}}$ for the subset of arithmetic weights. We sometimes write $\mathfrak{X}^d$ instead of $\mathfrak{X}$, d being the dimension of $\mathfrak{X}$, to emphasize the dimension of $\mathfrak{X}$.

For any rigid space $\mathfrak{U}$ we denote by $A(\mathfrak{U})$ the ring of analytic function on $\mathfrak{U}$. For our purposes a p-adic family of automorphic forms is a character

$$\lambda\colon\ \mathcal{R}_{S,p} \to A(\mathfrak{U}),$$

where $\mathfrak{U}$ is an irreducible rigid space over $\mathfrak{X}$ of positive dimension with projection map denoted w and such that there is a Zariski dense set of points $\Sigma \subset \mathfrak{U}(\overline{\mathbb{Q}}_p)^{\mathrm{alg}} := \{y \in \mathfrak{U}(\overline{\mathbb{Q}}_p) \mid \mathrm{w}(y) \in \mathfrak{X}^{\mathrm{alg}}(\overline{\mathbb{Q}}_p)\}$ with the property that for any $y \in \Sigma$ the compositum λ_y of λ with the evalutation map[3] at y is a normalized p-stabilization $\lambda_\pi^\dagger$ of an automorphic representation π of normalized weight $\mathrm{w}(y)^{\mathrm{alg}}$.

Definition 2.3.1 (strong form). A p-adic family of automorphic representations is a linear functional

$$I: \ \mathcal{C}_c^\infty(G(\boldsymbol{A}_S), \mathbb{Z}_p) \otimes \mathcal{R}_{S,p} \to A(\mathfrak{U})$$

such that $\lambda_I := I|_{\mathcal{R}_{S,p}}$ is a p-adic family in the weak sense and such that for all $y \in \Sigma$ as above, the compositum I_y of I with the evalutation map at y is the p-stabilized distribution $I_\pi^\dagger$ of an automorphic representation π of normalized weight $\mathrm{w}(y)^{\mathrm{alg}}$.

Note that these definitions, and hence all subsequent discussion, are relative to some fixed set S of primes containing the prime p.

A fundamental question in the area is whether an individual p-stabilization $I_\pi^\dagger$ is a member of a p-adic family. A positive answer to this question has been given by Hida in the ordinary case. Using the techniques of [Hi04], it can be shown that any ordinary p-adic *almost cuspidal*[4] eigenform is the member of a (strong) p-adic family of almost cuspidal eigenforms of dimension d. We say that an holomorphic modular form for $G_{a,b}$ is almost cuspidal if its constant terms along the parabolic subgroup $P_{a-1,b-1}$ are cuspidal. Using the techniques of [SU06], this can be generalized to forms with slope s satisfying $s_b = s_{b+1}$ (i.e., the semi-ordinary case, which means that $I_\pi^\dagger(u_0)$ is a p-adic unit where u_0 is the operator corresponding to the trace of the relative Frobenius [5] on the Shimura variety associated to $G_{a,b}$ in characteristic p. Note also that the following theorem is a special case of the results of [U06].

Theorem 2.3.2. *If I_0 is a non-critical cuspidal finite slope distribution of regular arithmetic weight w_0, then there exists $\mathfrak{U} \xrightarrow{\mathrm{w}} \mathfrak{X}$ of dimension d, $y_0 \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$, and a p-adic $\mathfrak{U}$-family I such that*

$$I_{y_0} = I_0 + I_1 + \cdots + I_s$$

with $I_1, \ldots, I_s$ irreducible character distributions of $C_c^\infty(\boldsymbol{A}_S) \otimes \mathcal{R}_{S,p}$ such that

$$I_i|_{\mathcal{R}_{S,p}} = I_0|_{\mathcal{R}_{S,p}} \quad \textit{for all } i = 1, \ldots, s.$$

We expect that a similar result must be true for general *overconvergent* modular forms. Using techniques from Kisin–Lai [KL], it is possible to construct such a deformation provided one only requires it to be of dimension one. We will use this technique for critical Eisenstein series.

[3] The map $A(\mathfrak{U}) \to \overline{\mathbb{Q}}_p$ given by $f \mapsto f(y)$.

[4] This terminology is non-standard.

[5] It corresponds to the d-tuple $(\underbrace{1, \ldots, 1}_{b}, \underbrace{p, \ldots, p}_{a})$.

3. Deformations of Eisenstein series

We keep to the notation of Sections 1 and 2. Recall that we have groups $G = G_{a,b}$ and $H = G_{a',b'}$ and $L = G \times \mathrm{Res}_{\mathcal{K}/\mathbb{Q}}\boldsymbol{G}_m$ a standard Levi subgroup of a parabolic P of H. In this section, we will consider specific p-adic families for the groups G and H. Keeping with our practice from Section 1, we will add a superscript $'$ when the notion is relative to H. For instance, I'_m means an Iwahori subgroup of $H(\mathbb{Q}_p)$.

3.1. Critical Eisenstein series. We now fix a cuspidal tempered representation π of $G(\boldsymbol{A})$ and an idele class character χ of $\boldsymbol{A}_{\mathcal{K}}^{\times}$ as in §1.4. We will assume that

$$\chi^{1+c} = |\cdot|_{\boldsymbol{A}_{\mathcal{K}}}^{2\kappa'} \tag{3.1.1}$$

and that the assumptions of Proposition 1.4.1 are satisfied along with

$$L(\pi, \chi^{-1}, \kappa' + 1/2) = 0. \tag{3.1.2}$$

To simplify matters, we will also assume that π and χ are unramified at primes above p. We let S be the set comprising the primes of ramification of π, χ, and G (and hence also of H) together with p. Let $m > 0$ be an integer. Then $\pi_p^{I_m} \neq 0$, and we choose $v_0 \in \pi_p^{I_m}$ a p-stabilization of π_p. We consider the section $\Phi_p^{\mathrm{crit}} \in I(\rho_p)$ defined for all $h \in H(\mathbb{Q}_p)$ by

$$\Phi_p^{\mathrm{crit}}(h) = \begin{cases} \chi(t)\pi_p(g).v_0 & \text{if } h = n.m(g,t)wk_0 \in P(\mathbb{Q}_p)wI'_m, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.3}$$

Here w is the Weyl element from (1.4.1).

For $s \in \mathbb{C}$ let $I(\rho_v, s) = \{\varphi_s = \varphi\delta^{s+1/2} \ ; \ \varphi \in I(\rho_v)\}$. The following lemma follows from a direct computation.

Lemma 3.1.1. *For each $s \in \mathbb{C}$ the section $\Phi_{p,s}^{\mathrm{crit}} = \Phi_p^{\mathrm{crit}}\delta^{s+1/2} \in I(\rho_p, s)$ is an eigenvector for the action of $\mathcal{U}'_p$. Moreover, if $(\alpha_1, \ldots, \alpha_d)$ is the ordering of Langlands parameters specifying the chosen p-stabilization v_0 of π_p then the ordering associated to $\Phi_{p,s_0}^{\mathrm{crit}}$ is given by*

$$(\alpha_1, \ldots, \alpha_b, \chi(\varpi)p^{\kappa'}, \chi(\varpi)p^{\kappa'+1}, \alpha_{b+1}, \ldots, \alpha_d),$$

and if the slope of v_0 is $(s_1, \ldots, s_d)$ then the slope of $\Phi_{p,s_0}^{\mathrm{crit}}$ is

$$(s_1, \ldots, s_b, 1, -1, s_{b+1}, \ldots, s_d).$$

In particular, it is critical[6].

[6] In contrast, the semi-ordinary p-stabilization obtained by taking $\Phi_p^{\mathrm{ord}} := M(\Phi_p^{\mathrm{crit},\vee}, -s_0)$, where $\Phi_p^{\mathrm{crit},\vee} \in I(\rho_p^{\vee})$ is defined analogously to Φ_p^{crit}, has slope $(s_1, \ldots, s_b, 0, 0, s_{b+1}, \ldots, s_d)$.

We consider the space V_0 generated by the Eisenstein series $E(\mathcal{F}_h; s_0, Z)$ associated to the sections $\Phi = \Phi_{\infty,i} \otimes \Phi_p^{\mathrm{crit}} \otimes \Phi^{p,\infty}$ with $\Phi^{p,\infty} = \otimes_{v \neq p,\infty} \Phi_v$ and $\Phi_v = \Phi_v^{\mathrm{sph}}$ if $v \notin S$. We let $V_1 \subset V_0$ be the subspace of Eisenstein series as above with the extra condition that $M(\Phi_v, s_0) = 0$ for all $v \in S \backslash \{p\}$ and let $E^{\mathrm{cr}}(\pi, \chi) := V_0/V_1$. This last space, a quotient of a space of almost cuspidal holomorphic automorphic forms for H of weight $\xi = (c_d, \ldots, c_{b+1}, \kappa - d/2 - 1; c_1, \ldots, c_b, \kappa + d/2 + 1)$, is acted on by $\mathcal{R}_{S,p} \otimes C_c^\infty(H(\boldsymbol{A}_S), \mathbb{Z}_p)$ and decomposes as

$$E^{\mathrm{cr}}(\pi, \chi) = \bigotimes_{v \in S \backslash \{p\}} \mathcal{L}(\pi_v, \chi_v, s_0)$$

with $\mathcal{L}(\pi_v, \chi_v, s_0)$ the Langlands quotient of $I(\rho_v, s_0)$. We denote by $I_{E^{\mathrm{cr}}(\pi,\chi)}$ the corresponding distribution of $\mathcal{R}_{S,p} \otimes C_c^\infty(H(\boldsymbol{A}_S), \mathbb{Z}_p)$.

For any finite place v of $\mathcal{K}$ and any representation Π_v of $\mathrm{GL}_n(\mathcal{K}_v)$, we denote by $\mathrm{rec}(\Pi_v)$ the n-dimensional representation of the Weil–Deligne group associated to Π_v by the local Langlands correspondence as established by Harris–Taylor [HT01]. Then we have

$$\mathrm{rec}(\mathrm{BC}(\mathcal{L}(\pi_v, \chi_v, s_0))) = \mathrm{rec}(\mathrm{BC}(\pi)_v) \oplus \chi_v \circ \mathrm{Art}_{\mathcal{K}_v}^{-1} \oplus \varepsilon^{-1}\chi_v \circ \mathrm{Art}_{\mathcal{K}_v}^{-1}$$

where $\mathrm{Art}_{\mathcal{K}_v}$ stands for the Artin reciprocity map sending a uniformizer to a geometric Frobenius and where ε denotes the cyclotomic character.

3.2. p-adic deformations. Let $\mathfrak{X}^{d+2}/\mathbb{Q}_p$ be the weight space for H. For any $w_0 = (c_1, \ldots, c_{d+2}) \in \mathfrak{X}^{d+2}$, we put

$$\mathfrak{X}_{w_0}^{d+2} = \{w = (e_1, \ldots, e_{d+2}) \in \mathfrak{X}^{d+2} \mid e_i - e_{i+1} = c_i - c_{i+1} \text{ for all } i \neq b+1\}.$$

This is clearly a two-dimensional closed subspace of $\mathfrak{X}^{d+2}$.

Theorem 3.2.1. *Suppose that* $w_0 = w_\xi = (c_1 - a - 1, \ldots, c_b - a - 1, \kappa + (b-a)/2, \kappa + (b-a)/2, c_{b+1} + b + 1, \ldots, c_d + b + 1)$. *There exist an affinoid* $\mathfrak{U}$ *sitting over* $\mathfrak{X}_{w_0}^{d+2}$, *a point* $y_0 \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$ *over* w_0, *and a two-dimensional family* $I : C_c^\infty(H(\boldsymbol{A}_S), \mathbb{Z}_p) \otimes \mathcal{R}_{S,p} \to A(\mathfrak{U})$ *such that*

$$I_{y_0} = I_{E^{\mathrm{cr}}(\pi,\chi)} + I_1 + \cdots + I_s$$

with the I_i*'s irreducible characters distributions satisfying*

$$I_i|_{\mathcal{R}_{S,p}} = I_{E^{\mathrm{cr}}(\pi,\chi)}|_{\mathcal{R}_{S,p}} \quad \textit{for all } i = 1, \ldots, s.$$

If $\pi_\infty = \pi_\tau^H$ *with* τ *regular, then this family extends to a* $d+2$*-dimensional family over* $\mathfrak{X}^{d+2}$.

Proof. We give just an idea of how this theorem is proved. The details will appear elsewhere. The proof does not require one to start from an Eisenstein series. The

techniques one uses to prove the first point of this theorem are similar to those used by Coleman, Kisin–Lai, and Kassaei. The deformations are constructed by studying the compact action of u_0 on the space of overconvergent modular forms for H obtained by multiplying one of the original critical Eisenstein series by powers of a characteristic zero lifting of powers of the Hasse invariant[7]. This requires that we first establish the rationality of (scalar multiples of) our critical Eisenstein series. The proof then employs the theory of the canonical subgroup as developed by various authors (Abbes–Mokrane, Kisin–Lai, Conrad). This provides a one-variable family. To obtain a two-variable family, one twists the one-variable family by anticyclotomic characters of p-power conductor. To prove the second point, one shows that the constructed curve sits in the eigenvarieties associated to H in [U06]. The regularity condition on τ should not be necessary in this special case. In general, however, it might be necessary to make sure that the 'classical' systems of Hecke eigenvalues occurring in the one variable family contribute only to the middle cohomology of the Shimura variety for H. □

The next lemma helps describe the restrictions $I|_{C_c^\infty(H(\mathbb{Q}_v))}$, $v \in S\backslash\{p\}$. Its proof will appear elsewhere.

Lemma 3.2.2. *Let π_0 be a unitary irreducible representation of $G(\mathbb{Q}_v)$ and χ_0 a unitary character of $\mathcal{K}_v^\times$. Let $J : C_c^\infty(H(\mathbb{Q}_v)) \to A(\mathfrak{U})$ be an analytic $\mathfrak{U}$-family of local character distributions such that*

$$J_{x_0}(f) = \mathrm{tr}(\mathcal{L}(\pi_0, \chi_0, s_0)(f)) + I_1(f) + \cdots + I_s(f)$$

where $I_1, \ldots, I_s$ are irreducible character distributions of $H(\mathbb{Q}_v)$. Assume J is generically irreducible. Then one of the two following cases holds:

(i) *There exist an analytic $\mathfrak{U}$-family of representations π of $G(\mathbb{Q}_v)$ and an analytic $\mathfrak{U}$-family of characters χ of $\mathcal{K}_v$ such that $J_x(f) = \mathrm{tr}(L(\pi_x, \chi_x, s_0)(f))$ for all $x \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$.*

(ii) *The place v is split. There exist an analytic $\mathfrak{U}$-family of representations π of $\mathrm{GL}_d(\mathbb{Q}_v)$ and two analytic $\mathfrak{U}$-families of characters μ and ν of $\mathbb{Q}_v$ with $\mu \neq \nu|\cdot|_v^{\pm 1}$ such that $J_x^H(f) = \mathrm{tr}((\mu_x \times \pi_x \times \nu_x)(f))$ for a Zariski dense set of points $x \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$ where $\mu_x \times \pi_x \times \nu_x$ is the irreducible induction $\mathrm{Ind}_{\boldsymbol{G}_m \times \mathrm{GL}_d \times \boldsymbol{G}_m}^{\mathrm{GL}_{d+2}} \mu_x \otimes \pi_x \otimes \nu_x$.*

4. Galois representations and applications to Selmer groups

4.1. Galois representations for automorphic representations. We begin with notation for the local theory.

[7] This is possible thanks to the theory of the arithmetic toroidal compactification of the Shimura variety associated to H by K. Fujiwara [Fu]

Let w be a finite place of $\mathcal{K}$ and $G_{\mathcal{K}_w}$ the absolute Galois group of the completion of $\mathcal{K}$ at w. We denote by $\mathrm{Frob}_w \in G_{\mathcal{K}_w}$ a geometric Frobenius element, $I_{\mathcal{K}_w} \subset G_{\mathcal{K}_w}$ the inertia subgroup, and $W_{\mathcal{K}_w} \subset G_{\mathcal{K}_w}$ the Weil subgroup.

Assume first that the residual characteristic of w is not p. To any finite-dimensional representation $R\colon G_{\mathcal{K}_w} \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$, one associates a Weil–Deligne representation $\mathrm{WD}(R) = (r, N)$ where $r\colon W_{\mathcal{K}_w} \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ is a representation and $N \in M_n(\overline{\mathbb{Q}}_p)$ is such that

$$R(\mathrm{Frob}_w^m \sigma) = r(\mathrm{Frob}_w^m \sigma) exp(t(\sigma)N)$$

where $t\colon I_{\mathcal{K}_w} \to \mathbb{Z}_p$ is defined by $\sigma(\sqrt[p^f]{\varpi_w}) = \zeta_{p^f}^{t(\sigma)} \cdot \sqrt[p^f]{\varpi_w}$ for a fixed choice of a compatible system $\{\zeta_{p^f}\}$ of p-power roots of unity and a uniformizer ϖ_w of $\mathcal{K}_w$. It is well-known that (r, N) is uniquely defined up to isomorphism.

If the residual characteristic of w is equal to p, one generally uses Fontaine's rings to study the p-adic representations of $G_{\mathcal{K}_w}$. If V is such a representation, one defines $D_?(V) = (V \otimes_{\mathbb{Q}_p} B_?)^{G_{\mathcal{K}_w}}$ with $? = \mathrm{dR}$, cris or st, where B_{dR}, B_{cris} and B_{st} are the usual rings of p-adic periods introduced by Fontaine. We write $D^i_{\mathrm{dR}}(V)$ for the i-th step of the Hodge filtration of $D_{\mathrm{dR}}(V)$. We adopt the geometric conventions for the Frobenius and the Hodge–Tate weights (so the Hodge–Tate weights of V are the jumps of the Hodge filtration of $D_{\mathrm{dR}}(V)$).

In both the local and global cases, we denote by ε_p the p-adic cyclotomic character and we write $V(n)$ for the n-th Tate twist of a Galois representation V.

Let now $\pi = \pi_f \otimes \pi_\infty$ be an automorphic representation of $G(\boldsymbol{A})$ such that $\pi_\infty = \pi_\tau^H$ for some $\tau = (c_d, \ldots, c_{b+1}; c_1, \ldots, c_b)$. Let $\boldsymbol{\kappa}_\tau = (\kappa_1, \ldots, \kappa_d)$ be the strictly increasing sequence of integers defined by

$$\kappa_{d-i+1} := c_i + d - i + \delta_i \quad \text{for all } i = 1, \ldots, d,$$

where $\delta_i = -a$ if $i \leq b$ and $\delta_i = b$ if $i \geq b+1$. Let S_π be the set of finite places of $\mathcal{K}$ above primes of ramification of π. The following conjecture[8] is expected to result from the stabilization of the trace formula for unitary groups.

Conjecture 4.1.1. There exists a finite extension L of $\overline{\mathbb{Q}}_p$ and a Galois representation

$$R_p(\pi)\colon G_{\mathcal{K}} \longrightarrow \mathrm{GL}_d(L)$$

satisfying the following properties:

1. $R_p(\pi)^\vee(1-d) \cong R_p(\pi)^c$.
2. $R_p(\pi)$ is unramified outside $S_\pi \cup \{\wp, \bar{\wp}\}$.
3. For each finite place w of $\mathcal{K}$ of residue characteristic prime to p, we have

$$\mathrm{WD}(R_p(\pi)|_{W_{K_w}}) \cong \mathrm{rec}(\mathrm{BC}(\pi)_w^\vee \otimes |\det|^{\frac{1-d}{2}})$$

[8] This conjecture is a theorem for unitary groups appearing in the works of Kottwitz, Clozel, Harris–Taylor, Yoshida-Taylor [HT01, TY06]

where rec is the reciprocity map given by the Local Langlands correspondence of Harris–Taylor [HT01] (using our identification of $\mathbb{C}_p$ with $\mathbb{C}$).

4. $R_p(\pi)|_{G_{\mathcal{K}_\wp}}$ is Hodge–Tate with Hodge–Tate weight given by $\boldsymbol{\kappa}_\tau$.

5. If π_p is unramified, then the eigenvalues of the Frobenius endomorphism of $D_{\mathrm{cris}}(R_p(\pi))$ are given by the Langlands parameters of π_p (again using the identification of $\mathbb{C}_p$ with $\mathbb{C}$).

Let χ_p be the Galois character of $G_{\mathcal{K}}$ associated to an idele class character χ as in §1.4 (i.e., such that $\chi_p(\mathrm{Frob}_w) = \chi(\varpi_w)$ if χ is unramified at w). We see in particular that (3) implies that

$$L^{\{p\}}(R_p(\pi) \otimes \chi_p, s) = L^{\{p\}}\left(\pi^\vee, \chi, s + \frac{1-d}{2}\right) \tag{4.1.1}$$

where $L^{\{p\}}$ means we have omitted the Euler factor at p and the L-function for the Galois representation is defined using the geometric Frobenius elements. Moreover, if χ also satisfies (3.1.1) then (4.1.1) implies

$$L^{\{p\}}(R_p(\pi) \otimes \chi_p, s) = L^{\{p\}}\left(\pi, \chi^{-1}, s + 2\kappa' + \frac{1-d}{2}\right). \tag{4.1.2}$$

4.2. Families of Galois representations. Let $\mathfrak{U}$ be a smooth connected affinoid variety defined over a p-adic field, and let G_F be the absolute Galois group of a number or ℓ-adic field F (ℓ may be equal to p). We call a pseudo-representation $T\colon G_F \to A^0(\mathfrak{U})$ an analytic family of Galois representations over $\mathfrak{U}$. For any reduced affinoid subdomain $\mathfrak{Z} \subset \mathfrak{U}$, we denote by $R^T_{\mathfrak{Z}}$ the semi-simple Galois representation (defined up to isomorphism) over a finite extension of the fraction ring $F(\mathfrak{Z})$ of $\mathfrak{Z}$ whose trace is the pseudo-representation $G_F \to A^0(\mathfrak{U}) \to A^0(\mathfrak{Z})$. We say that T is n-dimensional if $R^T_{\mathfrak{Z}}$ is. If $\mathcal{L}$ is a G_F-stable lattice of $R^T_{\mathfrak{U}}$ and $y \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$, then we denote by $R^{\mathcal{L}}_y$ the representation on the specialization $\mathcal{L} \otimes_{A_0(\mathfrak{U})} A(\mathfrak{U})/I_y$, where I_y is the ideal of analytic function on $\mathfrak{U}$ vanishing at y.

Assume F is a p-adic field. Let T be a family of representations of G_F of dimension d over an affinoid $\mathfrak{U}$. We denote by $\kappa_1, \ldots, \kappa_d \in A(\mathfrak{U})$ the Hodge–Tate–Sen weights of T. Let r be the dimension of the affinoid $\mathfrak{U}$. The family T is said to be of finite slope[9] if there exist

(i) $\varphi_1, \ldots, \varphi_d \in A^0(\mathfrak{U})$,

(ii) $\Sigma \subset \mathfrak{U}(\overline{\mathbb{Q}}_p)$,

(iii) a subset of $\{1, \ldots, d\}$ of r positive integers $i_1 \leq \cdots \leq i_r$,

such that

[9]This is "trianguline" in the terminology of Colmez.

(a) for all $y \in \Sigma$, we have the inequalities $\kappa_1(y) \leq \cdots \leq \kappa_d(y)$,

(b1) for all $i \notin \{i_1, \ldots, i_r\}$, $y \mapsto \kappa_{i+1}(y) - \kappa_i(y)$ is constant on $\mathfrak{U}$,

(b2) for all positive real numbers C, the subset Σ_C of points $y \in \Sigma$ such that $\kappa_{i_j}(y) - \kappa_{i_{j+1}}(y) > C$ for all $i = 1, \ldots, r$ is Zariski dense.

(c) for all $y \in \Sigma$, R_y^T is crystalline, and the eigenvalues of Frobenius on $D_{\mathrm{cris}}(R_y^T)$ are given by $\varphi_1(y)p^{\kappa_1(y)}, \ldots, \varphi_d(y)p^{\kappa_d(y)}$

Let $y_0 \in \mathfrak{U}(L)$ such that $(\kappa_1(y_0), \ldots, \kappa_d(y_0))$ is an increasing sequence of integers. According to a terminology of B. Mazur [M00], we say that T is a finite slope deformation of R_{y_0} of refinement $(\varphi_1(y_0)p^{k_1(y_0)}, \ldots, \varphi_d(y_0)p^{k_d(y_0)})$ and Hodge–Tate variation $(i_1, i_2 - i_1, \ldots, d - i_r)$.

Of course, there is a close link between p-stabilization and refinement. More precisely, we have the following easy lemma.

Lemma 4.2.1. *Assume Conjecture* 4.1.1. *Let π be a cuspidal representation which is tempered and unramified at p and of weight τ (i.e., $\pi_\infty = \pi_\tau^H$). Let $\mathfrak{U}$ be an affinoid sitting over $\mathfrak{X}$ and let I be a p-adic deformation of π with p-stabilization given by $(\alpha_1, \ldots, \alpha_d)$. Then there exists a p-adic deformation T_I over $\mathfrak{U}$ of $R_p(\pi)$ such that the restriction of T_I to $G_{\mathcal{K}_\wp}$ is a finite slope deformation of $\rho_\pi|_{G_{\mathcal{K}_\wp}}$ of refinement $(\varphi_1 p^{\kappa_1}, \ldots, \varphi_d p^{\kappa_d})$ with*

$$\varphi_i = \alpha_{d-i+1}^{-1} \cdot p^{-\kappa_i + (d-1)/2} \quad \text{for all } i = 1, \ldots, d, \tag{4.2.1}$$

where $\boldsymbol{\kappa}_\tau = (\kappa_1, \ldots, \kappa_d)$.

Proof. The existence of T_I follows from the theory of pseudo-representations. The asserted properties of the restriction of T_I to $G_{\mathcal{K}_\wp}$ follows from parts (4) and (5) of the Conjecture 4.1.1. The details are left to the reader. □

We recall the following useful result of Kisin.

Proposition 4.2.2. *Let $T : G_F \to A^0(\mathfrak{U})$ be a finite slope family as above. Let $\mathcal{L}$ be any G_F-stable free $A(\mathfrak{U})$-lattice of $R_{\mathfrak{U}}^T$. Let F_0 be the maximal unramified subfield of F. After shrinking $\mathfrak{U}$ around some fixed $y_0 \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$, the following holds:*

(i) *Let $1 \leq i \leq i_1$ be an integer. If $y \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$ is such that $\kappa_i(y) \in \mathbb{Z}$, then*

$$\mathrm{rank}_{L \otimes K_0} D_{\mathrm{cris}}(R_y^{\mathcal{L}})^{\phi = \varphi_i(y)p^{\kappa_i(y)}} \geq 1,$$

where $L = A(\mathfrak{U})/I_y$. Furthermore, there exists an integer N independent of y such that

$$D_{\mathrm{cris}}(R_y^{\mathcal{L}})^{\phi = \varphi_i(y)p^{\kappa_i(y)}} \hookrightarrow (R_y^{\mathcal{L}} \otimes B_{\mathrm{dR}}/t^{\kappa_i(y)+N} B_{\mathrm{dR}}^+)^{G_K}$$

for all $y \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$.

(ii) *Let* $Q_y(X) := \prod_{i=1}^{i_1}(X - \varphi_i(y)p^{\kappa_i(y)})$. *For any* y *such that* $\kappa_1(y) \in \mathbb{Z}$,

$$\mathrm{rank}_{L\otimes F_0} D_{\mathrm{cris}}(R_y^{\mathcal{L}})^{Q_y(\phi)=0} \geq i_1,$$

where $L = A(\mathfrak{U})/I_y$. *Furthermore, there exists an integer* N *independent of* y *such that*

$$D_{\mathrm{cris}}(R_y^{\mathcal{L}})^{Q_y(\phi)=0} \hookrightarrow (R_y^{\mathcal{L}} \otimes B_{\mathrm{dR}}/t^{\kappa_i(y)+N} B_{\mathrm{dR}}^+)^{G_K}$$

for all $y \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$.

Proof. The first part of the proposition, and hence the second part when $Q_y(X)$ has only simple roots, is a direct consequence of Corollary 5.3 of [Ki]. When $Q_y(X)$ has multiple roots a simple generalization of the argument of [Ki] does the job. We can also as suggested to us by M. Kisin apply (i) to the case $\mathfrak{V} := Sp(A(\mathfrak{U})[X]/(Q(X))$ at least when $Q(X)$ has only generic simple roots. □

We deduce from this proposition a few interesting consequences that we will use to construct elements in Selmer groups.

Lemma 4.2.3. *Let* T *be a finite slope* $\mathfrak{U}$*-family of representations of* G_F *as in Proposition* 4.2.2, *and let* $y \in \mathfrak{U}(L)$ *be such that* $R_y^T := L(1)^f \oplus L^e \oplus V^{ss}$ *for* V *a de Rham representation of* G_F. *We assume that*

(i) 1 *is a root of* $Q_y(X)$ *of order* e,

(ii) $D_{\mathrm{cris}}(V)^{\phi=1} = 0$.

Let $\mathcal{L}$ *be a free lattice such that we have an exact sequence*

$$0 \to V \to R_y^{\mathcal{L}} \to W \to 0,$$

then

(a) *any non trivial extension of* L *by* $L(1)$ *appearing as a subquotient of* W *is crystalline;*

(b) *if* E *is the inverse image of* W^{G_F} *by the projection map from* $R_y^{\mathcal{L}}$ *to* W, *then* E *is an extension of* W^{G_F} *by* V, *the class* $[E]$ *of which is contained in* $H_f^1(K, \mathrm{Hom}(W^{G_F}, V))$.

Proof. Since $D_{\mathrm{cris}}(W/W^{G_F})^{\phi=1}$ and $D_{\mathrm{cris}}(E)^{\phi=1}$ have rank at most $e - \dim W^{G_F}$ and $\dim W^{G_F}$, respectively, by hypothesis (ii), and since the rank of $D_{\mathrm{cris}}(R_y^{\mathcal{L}})^{\phi=1}$ is e by hypothesis (i) and Proposition 4.2.2, we deduce that the respective ranks of $D_{\mathrm{cris}}(W/W^{G_K})^{\phi=1}$ and $D_{\mathrm{cris}}(E)^{\phi=1}$ equal $e - \dim W^{G_K}$ and $\dim W^{G_K}$. From $D_{\mathrm{cris}}(V)^{\phi=1} = 0$ and $D_{\mathrm{cris}}(E)^{\phi=1}$ being of rank $\dim W^{G_K}$, we deduce the surjectivity in the following short exact sequence:

$$0 \to D_{\mathrm{cris}}(V) \to D_{\mathrm{cris}}(E) \to D_{\mathrm{cris}}(W^{G_K}) \to 0.$$

Exactness of this sequence means, by definition, that $[E] \in H^1_f(K, \mathrm{Hom}(W^{G_K}, V))$ and (b) is proved. The proof of (a) follows similarly using $\mathrm{rank} D_{\mathrm{cris}}(W/W^{G_K})^{\phi=1} = e - \dim W^{G_K}$. The details are left to the reader. □

The following lemma will be used in the last section of this paper.

Lemma 4.2.4. *Let K be a p-adic field and let R_0 be a de Rham representation of G_K over a finite extension L of $\mathbb{Q}_p$. Let $\mathfrak{U}$ be an affinoid and $T: G_K \to A(\mathfrak{U})$ a finite slope deformation of the character representation $T_0 = 1 + \mathrm{tr}(R_0)$ of refinement $\varphi_1, \ldots, \varphi_{n+1}$ and Hodge–Tate weight variation $(i_1, i_2 - i_1, \ldots, n + 1 - i_r)$ Let $\mathcal{L}$ be a free G_K-stable $A(\mathfrak{U})$-lattice. We assume the following hypotheses are satisfied.*

(i) $\varphi_i \neq 1$ *if* $i \leq i_1$.

(ii) *There exists* $y \in \mathfrak{U}(L)$ *such that* $T_y = T_0$ *and* $k_i(y) > 0$ *for* $i > i_1$.

(iii) *The representation* $R^{\mathcal{L}}_{\mathfrak{U}}$ *is an extension of the form*

$$0 \to A(\mathfrak{U}) \to R^{\mathcal{L}}_{\mathfrak{U}} \to S_{\mathfrak{U}} \to 0$$

with trivial action of G_K *on* $A(\mathfrak{U})$.

Then the rank over $L \otimes K$ *of* $\mathrm{gr}^0 D_{\mathrm{dR}}(R^{\mathcal{L}}_y) = (R^{\mathcal{L}}_y \otimes \mathbb{C}_p)^{G_K}$ *is one more than the rank of* $\mathrm{gr}^0 D_{\mathrm{dR}}(R_0)$.

Proof. We have to prove that the Sen operator determining the action of a finite index subgroup of G_K on $R^{\mathcal{L}}_y \otimes \mathbb{C}_p$ has the eigenvalue 0 with multiplicity $1 + h_0$ with $h_0 := \mathrm{rank}\ \mathrm{gr}^0 D_{\mathrm{dR}}(R_0)$. Equivalently, we need to show that the order of vanishing at 0 of the minimal polynomial of the Sen operator of $R^{\mathcal{L}}_y$ is one. By hypothesis (ii), it is easy to see that it is therefore sufficient to show the same statement for $R^{\mathcal{L}}_z$ for any z sufficiently closed to y and such that

(a) $k_i(z) = k_i(y)$ if $i \leq i_1$,

(b) $k_i(z) > C$ if $i > i_1$,

where C is any arbitrary large constant (we know that we can approach y by such points by the axioms of a finite slope deformation). We now prove the result for z satisfying (a) and (b).

After (if necessary) replacing $\mathfrak{U}$ by a sufficiently small neighborhood of y, we know by Proposition 4.2.2 that

$$D_{\mathrm{cris}}(R^{\mathcal{L}}_z)^{Q_z(\phi)=0} \otimes K \hookrightarrow (R^{\mathcal{L}}_z \otimes B_{\mathrm{dR}}/t^{k_{i_1}(z)+N} B^+_{\mathrm{dR}})^{G_K}.$$

If $C > N + k_{i_1}(y)$, we therefore have that if z satisfies (a) and (b) then the image of $D_{\mathrm{cris}}(R^{\mathcal{L}}_z)^{Q_z(\phi)=0} \otimes K \cap D^0_{\mathrm{dR}}(R^{\mathcal{L}}_z)$ in $\mathrm{gr}^0(D_{\mathrm{dR}}(R^{\mathcal{L}}_z))$ is of rank h_0. On the other hand, by our hypothesis (iii), we have an exact sequence

$$0 \to L \to R^{\mathcal{L}}_z \to S_z \to 0$$

and therefore $\mathrm{gr}^0(D_{\mathrm{dR}}(R_z^{\mathcal{L}}))$ contains also the non trivial image of $D_{\mathrm{cris}}(L)$ on which the action of ϕ is given by the eigenvalue 1. By hypothesis (i) we may assume that $Q_z(1) \neq 0$ for z sufficiently close to y and therefore the images of $D_{\mathrm{cris}}(L)$ and $D_{\mathrm{cris}}(R_z^{\mathcal{L}})^{Q_z(\phi)=0} \otimes K \cap D^0_{\mathrm{dR}}(R_z^{\mathcal{L}})$ in $\mathrm{gr}^0(D_{\mathrm{dR}}(R_z^{\mathcal{L}}))$ are disjoint and hence $\mathrm{gr}^0(D_{\mathrm{dR}}(R_z^{\mathcal{L}}))$ has rank $1 + h_0$. □

4.3. Deformations of some reducible Galois representations and Selmer groups. Let χ and π be as in §§1.4 and 3.1. Let S be a finite set of places of $\mathcal{K}$ containing $\wp, \wp^c$ and the primes of ramification of $\mathrm{BC}(\pi)$ and χ. Assuming $L(\pi, \chi^{-1}, \kappa' + 1/2) = 0$, we have constructed in §§1 and 3 an Eisenstein representation $E^{\mathrm{cr}}(\pi, \chi)$ whose S-primitive L-function is given by

$$L^S(E^{\mathrm{cr}}(\pi, \chi), s) = L^S(\pi, s)L^S(\chi, s - \kappa' - 1/2)L^S((\chi^c)^{-1}, s + \kappa' + 1/2).$$

Therefore the Galois representation associated to our Eisenstein representation is:

$$R_p(\pi)(-1) \oplus \chi_p^{-1}\varepsilon^{1+\kappa'-d/2} \oplus \chi_p^c\varepsilon^{\kappa'-d/2}.$$

Assume now that χ satisfies (3.1.1). We consider the Galois representation

$$R := R_p(\pi) \otimes \chi_p\varepsilon^{d/2-\kappa'}.$$

It satisfies

$$R^c \cong R^\vee(1) \tag{4.3.1}$$

and we therefore have the functional equation

$$L(R, -s) = \varepsilon(R, s)L(R, s),$$

and $s = 0$ is the central value for $L(R, s)$. By (4.1.2), $L^S(R, 0) = L^S(\pi, \chi^{-1}, \kappa' + 1/2)$. Note that the conditions on the weights of χ and π at the beginning of Section 1.4 imply that R does not have the Hodge–Tate weights 0 and -1. It can be seen that any (automorphic) irreducible Galois representation of $G_{\mathcal{K}}$ with regular Hodge–Tate weights having no Hodge–Tate weights equal to 0 and -1 and satisfying the condition (4.3.1) should be obtained in this way. Although it is not necessary, we will assume that R is irreducible.

The following result is suggested by the Bloch–Kato conjectures.

Theorem 4.3.1. *Assume Conjecture* 4.1.1 *for unitary groups in $d + 2$ variables. Assume π is tempered and that π and χ are unramified at primes above p. Assume also that $R_p(\pi)$ is irreducible. Then, if $L(R, 0) = 0$, we have*

$$\mathrm{rank}\ H^1_f(\mathcal{K}, R^\vee(1)) \geq 1.$$

Here $H^1_f(\mathcal{K}, R^\vee(1))$ is the Bloch–Kato Selmer group associated to the p-adic representation $R^\vee(1)$; for a definition see [BK] or [FP].

Proof. The proof of this theorem runs along the same lines as that of Theorem 4.1.4 in [SU06].

We first choose a non-critical p-stabilization $(\alpha_1, \dots, \alpha_d)$ of π and denote by $(\varphi_1, \dots, \varphi_d)$ the corresponding refinement of $R_p(\pi)$ (given by (4.2.1)). Recall that we write τ for the weight of π_∞ and ξ for the weight of the Eisenstein series. Since we assume the existence of Galois representations for cuspidal representations of the unitary group $G_{a+1,b+1}$, by Theorem 3.2.1 there exists a two-dimensional affinoid subdomain $\mathfrak{U}$ sitting over a closed subspace of $\mathfrak{X}^{d+2}_{w_\xi}$, a point $y_0 \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$ over $w_0 = w_\xi$, and a $\mathfrak{U}$-family T of Galois representations such that the specialization of T at y_0 is the pseudo-representation associated to $R_p(\pi)(-1) \oplus \chi_p^{-1}\varepsilon^{\kappa'-d/2} \oplus \chi_p^{-1}\varepsilon^{\kappa'-d/2-1}$ and such that the restriction of T to $G_{\mathcal{K}_\wp}$ is of refinement

$$(p\varphi_1, \dots, p\varphi_a, \chi(\varpi)^{-1}p^{d/2-\kappa'+1}, \chi(\varpi)^{-1}p^{d/2-\kappa'}, p\varphi_{a+1}, \dots, p\varphi_d)$$

and Hodge–Tate variation $(a+1, b+1)$ (i.e. $r = 1$ and $i_1 = a+1$). Furthermore, from Lemma 3.2.2 and property (3) of the Conjecture 4.1.1, for all finite places w of $\mathcal{K}$ prime to p,

$$R^T_{\mathfrak{U}}|_{G_{\mathcal{K}_w}} \cong \mu_1 \oplus R_w \oplus \mu_2 \tag{4.3.1}$$

where μ_1, μ_2 are two $A(\mathfrak{U})$-valued characters of $G_{\mathcal{K}_w}$ specializing to $\chi_p^{-1}\varepsilon^{\kappa'-d/2}|_{G_{\mathcal{K}_w}}$ and $\chi_p^{-1}\varepsilon^{\kappa'-d/2+1}|_{G_{\mathcal{K}_w}}$ at the point y and R_w is a d-dimensional representation specializing to $R_p(\pi)|_{G_{\mathcal{K}_w}}$ at y.

We consider the normalized deformation $\tilde{R}_{\mathfrak{U}} := R^T_{\mathfrak{U}} \otimes \chi_p\varepsilon_p^{d/2-\kappa'+1}$. We have $\tilde{R}^\vee_{\mathfrak{U}}(1) = \tilde{R}^c_{\mathfrak{U}}$, and the semi-simplified specialization of $\tilde{R}_{\mathfrak{U}}$ at $y \in \mathfrak{U}(L)$ is given by $\tilde{R}_{\mathfrak{U},y} = L \oplus L(1) \oplus R$. The restriction of $\tilde{R}_{\mathfrak{U}}$ to $G_{\mathcal{K}_\wp}$ is a deformation of $\tilde{R}_{\mathfrak{U},y}|_{G_{\mathcal{K}_\wp}}$ of refinement $(\beta_1, \dots, \beta_a, 1, p^{-1}, \beta_{a+1}, \dots, \beta_d)$ with $\beta_i = \varphi_i\chi(\varpi)p^{\kappa'-d/2}$ and of Hodge variation $(a+1, b+1)$. We deduce from this that the restriction of $\tilde{R}_{\mathfrak{U}}$ to the decomposition subgroup $G_{\mathcal{K}_{\wp^c}}$ is a deformation of $\tilde{R}_{\mathfrak{U},y}|_{G_{\mathcal{K}_{\wp^c}}}$ of refinement $(p^{-1}\beta_d^{-1}, \dots, p^{-1}\beta_{a+1}^{-1}, 1, p^{-1}, p^{-1}\beta_a^{-1}, \dots, p^{-1}\beta_1^{-1})$ and of Hodge variation $(b+1, a+1)$.

We claim that $\mathrm{tr}(\tilde{R}^{ss}_{\mathfrak{U}})$ is not of the form $T' + T''$ where T' and T'' are two pseudo-representations. Were this the case, then they would have to satisfy $T'_y(g) = 1+\varepsilon_p(g)$ and $T''(g) = \mathrm{tr}(R(g))$ for all $g \in G_{\mathcal{K}}$. Assume this is so, and let us show we get a contradiction. First we show that the restriction to $G_{\mathcal{K}_\wp}$ of the representation R' associated to T' would be irreducible. By Proposition 4.2.2 the specialization of R' at any arithmetic point y' such that $s = \kappa_{b+2}(y') - \kappa_{b+1}(y') > 1$ would be a crystalline representation of Hodge–Tate weights $(0, s)$ and slopes $(1, s-1)$ and is therefore irreducible. The same statement holds for the restriction to $G_{\mathcal{K}_{\wp^c}}$. Moreover, the restriction of R' to $G_{\mathcal{K}_w}$ for $w \nmid p$ is a split sum of two characters by (4.3.1). Then exactly as in [SU06, Thms 4.2.7 or 4.3.4] we would deduce that there is an non-trivial extension class in $H^1_f(\mathcal{K}, \overline{\mathbb{Q}}_p(1))$; but we know that this group is trivial since the rank of the units in $\mathcal{K}$ is 0.

From the above discussion we deduce that $\tilde{R}_{\mathfrak{U}}$ is irreducible. Let $g \in G_{\mathcal{K}}$ be such that one of the eigenvalues, say α_0, of $\tilde{R}(g)$ is distinct from 1 and $\varepsilon_p(g)$ and choose α in some finite normal extension $A(\mathfrak{V})$ of $A(\mathfrak{U})$ such that $\alpha(z) = \alpha_0$ for some $z \in \mathfrak{V}(\overline{\mathbb{Q}}_p)$ above y. We take v in the representation space of $\tilde{R}'_{\mathfrak{V}} := \tilde{R}'_{\mathfrak{U}} \otimes_{A(\mathfrak{U})} A(\mathfrak{V})$ such that $g.v = \alpha.v$. We then consider the $A(\mathfrak{V})$-lattice $\mathcal{L}$ of $\tilde{R}'_{\mathfrak{V}}$ generated by $g.v$ over $A(\mathfrak{V})$ as g runs through $G_{\mathcal{K}}$. After possibly shrinking $\mathfrak{V}$ around z, we can assume $\mathcal{L}$ is free. By construction, $\mathcal{L}_z$ has a unique irreducible quotient, and this quotient is isomorphic to R. We therefore have an exact sequence of $G_{\mathcal{K}}$-representations

$$0 \to W \to R_z^{\mathcal{L}} \to R \to 0$$

with $W^{ss} \cong L \oplus L(1)$, L being the residue field of z. We first note that by (4.3.1) the restriction of W to $G_{\mathcal{K}_w}$, $w \nmid p$, is split. Moreover, we know by the application of Proposition 4.2.2 that $D_{\mathrm{cris}}(R_z^{\mathcal{L}})^{\phi=1}$ is non zero. Since 1 is not a root of Frobenius for R since $L(\mathrm{BC}(\pi)_w, \chi_w, 1/2)^{-1} \neq 0$ at $w|p$ by [JS, sect. 2.5], we deduce $D_{\mathrm{cris}}(W|_{G_{\mathcal{K}_w}})^{\phi=1}$ has rank 1 for all $w|p$. This shows that W is not a non-trivial extension of L by $L(1)$ since this extension would belong to $H^1_f(\mathcal{K}, L(1)) = 0$ (same argument as in [SU06, 4.3.4]).

Therefore L_z contains the trivial representation L and we can take $E := \mathcal{L}_z/L$. This gives a non-trivial extension:

$$0 \to R^\vee(1) \to E^\vee(1) \to L \to 0.$$

It follows from lemma 4.2.3, that $\mathrm{Res}_{\mathcal{K}_w}([E^\vee(1)]) \in H^1_f(\mathcal{K}_w, R^\vee(1))$ for $w|p$. Note that we again use the fact that 1 is not a root of the Frobenius for $R^\vee(1) \cong R^c$ as this is an hypothesis of the quoted lemma. If $w \nmid p$, $\mathrm{Res}_{\mathcal{K}_w}([E^\vee(1)]) \in H^1_f(\mathcal{K}_w, R^\vee(1))$ follows from (4.3.1). This ends the proof of the theorem. □

5. Higher order vanishing and higher rank Selmer groups

5.1. Higher order of vanishing. In this section, we assume, as in Theorem 4.3.1, that

$$L(\pi, \chi^{-1}, 1/2 + \kappa') = L(R, 0) = 0.$$

Since we are assuming (4.3.1), the primitive L-function of the Eisenstein representation $E^{\mathrm{cr}}(\pi, \chi)$ twisted by χ^{-1} is

$$L(E^{\mathrm{cr}}(\pi, \chi), \chi^{-1}, s) = L(\pi, \chi^{-1}, s)\zeta_{\mathcal{K}}(s - \kappa' - 1/2)\zeta_{\mathcal{K}}(s - \kappa' + 1/2).$$

Therefore the order of vanishing of $L^S(E^{\mathrm{cr}}(\pi, \chi), \chi^{-1}, s)$ at $s = s_0 = \kappa' + 1/2$ is one less than the order of vanishing of $L(\pi, \chi^{-1}, s)$ at $s = s_0$, because $\zeta_{\mathcal{K}}(0) \neq 0$ and $\zeta_{\mathcal{F}}(s)$ has a simple pole at $s = 1$. This remark is the starting point of a method of constructing a higher rank subspace in the Selmer group $H^1_f(\mathcal{K}, R^\vee(1))$ when such

a space is predicted by the Bloch–Kato conjecture (i.e., when $L(R, s)$ vanishes to higher order at $s = 0$). In this final section of this paper, we deal with the case of even order vanishing. More precisely, we will sketch a proof of the following theorem.

Theorem 5.1.1. *Let π and χ as in Theorem* 4.3.1. *We assume Conjecture* 4.1.1 *for $G_{a+2,b+2}$. If $L(R, s)$ vanishes to even order at $s = 0$, then*

$$\text{rank } H^1_f(\mathcal{K}, R^\vee(1)) \geq 2.$$

5.2. Sketch of proof. Since $L(R, s)$ vanishes to even order at $s = 0$, $\varepsilon(\pi, \chi^{-1}, 1/2+\kappa') = \varepsilon(R, 0) = 1$. This implies, by the remark at the beginning of this section, that

$$\varepsilon(R \oplus \varepsilon_p \oplus 1) = -\varepsilon(R, 0) = -1. \tag{5.2.1}$$

Let $\mathfrak{U}$ above $\mathfrak{X}^{d+2}$ and Σ be as in Theorem 3.2.1 and let $\tilde{R}_{\mathfrak{U}}$ be as in the proof of Theorem 4.3.1. By (5.2.1), for each $z \in \mathfrak{U}(\overline{\mathbb{Q}}_p)$, $\varepsilon(\tilde{R}_z, 0) = -1$. In particular, for each arithmetic point $z \in \Sigma^{\text{reg}}$, the subset of elements $z \in \Sigma$ with $\mathrm{w}(z) = (w_1, \ldots, w_{d+2})$ satisfying the regularity condition $w_{b+1} \geq \kappa + (b-a)/2 + 1$ and $w_{b+2} \leq \kappa + (b-a)/2 + 1$ (the analog of (1.4.1) with d replaced by $d+2$), we have $L(\pi(z), \chi^{-1}, 1/2+\kappa') = 0$ where $\pi(z)$ is the (holomorphic) cuspidal automorphic representation of $G_{a+1,b+1}(\boldsymbol{A})$ associated to z. We can therefore apply Proposition 1.4.1 and Theorem 3.2.1 with $\pi(z)$ and $G_{a+1,b+1}$ in place of π and $G_{a,b}$ and then repeat the argument of Theorem 4.3.1. Let ξ_z the weight of the Eisenstein series representation $E^{\text{cr}}(\pi(z), \chi)$. For each $z \in \Sigma$, there exists $\mathfrak{U}_z$ above $\mathfrak{X}^{d+4}_{w_{\xi_z}}$, a point $y_z \in \mathfrak{U}_z(\overline{\mathbb{Q}}_p)$ over w_{ξ_z}, and a pseudo-representation $T_z \colon G_{\mathcal{K}} \to A(\mathfrak{U}_z)$ as in the proof of Theorem 4.3.1.

Let w_1 be the arithmetic weight of $\mathfrak{X}^{d+4}$ defined by $w_1 := (c_1 - a - 2, \ldots, c_b - a - 2, \kappa + \frac{b-a}{2} - 1, \kappa + \frac{b-a}{2}, \kappa + \frac{b-a}{2}, \kappa + \frac{b-a}{2} + 1, c_{b+1} + b + 2, \ldots, c_d + b + 2)$. Let $\mathfrak{Y} \subset \mathfrak{X}^{d+4}$ be the set of weight $w = (e_1, \ldots, e_{d+4})$ such that $e_i = e_{i+1}$ for $i \neq b+1, b+2, b+3, d+4$. This is a 4-dimensional subspace of $\mathfrak{X}^{d+4}$. One can show there exists a 4-dimensional affinoid $\mathfrak{V}$ sitting over $\mathfrak{Y}_{w_1} := w_1 + \mathfrak{Y}$, containing $\mathfrak{U}_z$ for each $z \in \Sigma^{\text{reg}}$, and admitting a $\mathfrak{V}$-family of automorphic representations interpolating the $\mathfrak{U}_z$-families. In other words, the $\mathfrak{U}_z$-families fit together into a 4-dimensional family.

Let $S \colon G_{\mathcal{K}} \to A(\mathfrak{V})$ be the Galois deformation associated to the above $\mathfrak{V}$-family. It is a deformation of

$$\begin{aligned} S_{y_1} = \mathrm{tr}(R_p(\pi)(-2)) &+ \chi_p^{-1}\varepsilon_p^{\kappa'-d/2-1} + \chi_p^{-1}\varepsilon_p^{\kappa'-d/2-1} \\ &+ \chi_p^{-1}\varepsilon_p^{\kappa'-d/2-2} + \chi_p^{-1}\varepsilon_p^{\kappa'-d/2-2}. \end{aligned}$$

for some point $y_1 \in \mathfrak{V}(L)$ sitting over w_1. We consider the normalization defined by $\tilde{\tilde{R}}_{\mathfrak{V}} := R^S_{\mathfrak{V}} \otimes \chi_p \varepsilon_p^{d/2-\kappa'+2}$. Then, $\tilde{\tilde{R}}_{\mathfrak{V}}$ is a deformation of $\mathrm{tr}(R) + 1 + 1 + \varepsilon_p + \varepsilon_p$.

It is also a deformation of $\mathrm{tr}(R_z) + 1 + \varepsilon_p$ for all $z \in \Sigma^{\mathrm{reg}}$ where we have written $R_z := R_p(\pi(z)) \otimes \chi_p \varepsilon_p^{d/2-\kappa'+1}$. From the construction, it follows also that $\tilde{\tilde{R}}_{\mathfrak{V}}|_{G_{\mathcal{K}_\wp}}$ is a finite slope deformation of refinement $(\beta_1, \ldots, \beta_a, 1, 1, p^{-1}, p^{-1}, \beta_{a+1}, \ldots, \beta_d)$ and Hodge variation type $(a+1, 1, 1, b+1)$, and similarly for the restriction of $\tilde{\tilde{R}}_{\mathfrak{V}}$ to $G_{\mathcal{K}_{\wp^c}}$.

As in Theorem 4.3.1, we consider a lattice $\mathcal{L} \subset \tilde{\tilde{R}}_{\mathfrak{V}}$ such that the specialization $R_{y_1}^{\mathcal{L}}$ has a unique quotient isomorphic to R. In particular, this implies that $R_z^{\mathcal{L}}$ has a unique quotient isomorphic to R_z. Moreover, from the proof of Theorem 4.3.1 applied to $\pi(z)$, we see that $R_z^{\mathcal{L}}$ contains the trivial representation as a unique subrepresentation and has a quotient defining an extension E_z whose class belongs to $H_f^1(\mathcal{K}, R_z^\vee(1))$. In what follows, we will assume for simplicity that $\mathcal{L}$ is free although this might not be the case in general. However, it would not be difficult – although a bit cumbersome – to put ourselves in such a situation with a 'localization' argument similar to the one used in [SU06, §4.3.2].

Let $\mathfrak{U} := \mathfrak{V} \times_{\mathfrak{Y}_{w_1}} \mathfrak{X}_{w_1}^{d+4}$. This is the Zariski closure of Σ^{reg}. It follows from the discussion above that $R_{\mathfrak{V}}^{\mathcal{L}} \otimes A(\mathfrak{U})$ contains the trivial representation and that the quotient $\tilde{E}$ by the latter is an extension

$$0 \to A(\mathfrak{U}).\varepsilon_p \to \tilde{E} \to \tilde{R} \to 0$$

where $\tilde{R}$ is the deformation of $\mathrm{tr}(R) + 1 + \varepsilon_p$ having a unique quotient isomorphic to R that appeared in the proof of Theorem 4.3.1. Then the restriction of $\tilde{E}$ to $G_{\mathcal{K}_\wp}$ is a finite slope deformation of refinement $(\beta_1, \ldots, \beta_a, 1, p^{-1}, p^{-1}, \beta_{a+1}, \ldots, \beta_d)$ and Hodge variation type $(a+1, b+2)$, and similarly for the restriction of $\tilde{E}$ to $G_{\mathcal{K}_{\wp^c}}$.

We now study the specialization $\tilde{E}_{y_1}$. It has a unique quotient isomorphic to R and has semi-simplification $L \oplus L(1) \oplus L(1) \oplus R$. We first remark that the trivial representation has to be a subrepresentation of $\tilde{E}_{y_1}$, for otherwise the latter would contain a non-trivial extension of L by $L(1)$. This extension would be unramified outside p and crystalline at p by another application of Proposition 4.2.2 and therefore would give a non-trivial element in $H_f^1(\mathcal{K}, L(1))$.

Quotienting $\tilde{E}_{y_1}$ by this trivial representation, we get an extension E_1. We will now prove that E_1 contains $L(1) \oplus L(1)$. Otherwise, it will contain a non-trivial extension of $L(1)$ by $L(1)$. It is easy to see that this extension would be unramified outside p. It would also be Hodge–Tate by Lemma 4.2.4 applied to $E_1 \otimes L(-1)$ with $R_0 = R(-1) \oplus L$. Such a non-trivial extension does not exist.

We deduce that $E_1^\vee(1)$ is an extension of the form

$$0 \to R^\vee(1) \to E_1^\vee(1) \xrightarrow{f} V \to 0$$

with V a L-vector space of dimension 2 with trivial action of Galois. We deduce that we have an exact sequence

$$0 \to H^0(\mathcal{K}, R^\vee(1)) \to H^0(\mathcal{K}, E_1^\vee(1)) \to V \xrightarrow{\delta} H^1(\mathcal{K}, R^\vee(1).$$

We have $H^0(\mathcal{K}, E_1^\vee(1)) = 0$, for otherwise $E_1^\vee(1)$ contains the trivial representation, but $R^\vee(1)$ is the only subrepresentation of $E^\vee(1)$ since R is the only quotient of E_1 and $R^\vee(1) \cong R^c$ does not contain the trivial representation by hypothesis. Thus δ is injective. We can show that its image is contained in $H_f^1(\mathcal{K}, R^\vee(1))$ using Lemma 4.2.3 just as we proved this for the class $[E^\vee(1)]$ in the proof of Theorem 4.3.1. Since V is of dimension 2, this proves the theorem. □

References

[BC04] Bellaïche, J., and Chenevier, G., Formes non tempres pour U(3) et conjectures de Bloch-Kato. *Ann. Sci. École Norm. Sup.* (4) **37** (4) (2004), 611–662.

[BR92] Blasius, D., and Rogawski, J., Tate classes and arithmetic quotients of the two-ball. In *The zeta functions of Picard modular surfaces*, Université de Montréal, Centre de Recherches Mathématiques, Montreal, QC, 1992, 421–444.

[BK] Bloch, S., and Kato, K., L-functions and Tamagawa numbers of motives. In *The Grothendieck Festschrift*, Vol. I, Progr. Math. 86, Birkhäuser, Boston, MA, 1990, 333–400.

[FP] Fontaine, J.-M., Perrin-Riou, B., Autour des conjectures de Bloch et Kato: cohomologie galoisienne et valeurs de fonctions L. In *Motives* (Seattle, WA, 1991), Proc. Sympos. Pure Math. 55, Part 1, Amer. Math. Soc., Providence, 1994, 599–706.

[Fu] Fujiwara, K., Arithmetic compactifications of Shimura varieties. Preprint 1991.

[HL04] Harris, M., and Labesse, J.-P., Conditional base change for unitary groups. *Asian J. Math.* **8** (4) (2004), 653–683.

[HT01] Harris, M., Taylor, R., *The geometry and cohomology of some simple Shimura varieties*. Ann. of Math. Stud. 151, Princeton University Press, Princeton, NJ, 2001.

[Hi04] Hida, H., *p-adic automorphic forms on Shimura varieties*. Springer Monogr. Math., Springer-Verlag, New York 2004.

[JS] Jacquet, H., Shalika, J., On Euler products and classification of automorphic representations I. *Amer. J. Math.* **103** (3 (1981), 499–558.

[Ka04] Kato, K., p-adic Hodge theory and values of zeta functions of modular forms. Cohomologies p-adiques et applications arithmétiques. III. *Astérisque* **295** (2004), 117–290.

[Ki] Kisin, M., Overconvergent modular forms and the Fontaine-Mazur conjecture. *Invent. Math.* **153** (2) (2003), 373–454.

[KL] Kisin, M., Lai, K. F., Overconvergent Hilbert modular forms. *Amer. J. Math.* **127** (4) (2005), 735–783.

[M00] Mazur, B., The theme of p-adic variation. In *Mathematics: frontiers and perspectives*, Amer. Math. Soc., Providence, RI, 2000, 433–459.

[Sh] Shahidi, F., Functional equation satisfied by certain L-functions. *Compositio Math.* **37** (2) (1978), 171–207.

[SU02] Skinner, C., and Urban, E., Sur les déformations p-adiques des formes de Saito-Kurokawa. *C. R. Math. Acad. Sci. Paris* **335** (7) (2002), 581–586.

[SU06] Skinner, C., and Urban, E., Sur les déformations p-adiques de certaines représentations automorphes. *J. Inst. Math. Jussieu*, to appear.

[SU-MC] Skinner, C., and Urban, E., The Iwasawa main conjectures for GL_2. In preparation.

[TY06] Taylor, R., and Yoshida, T., Compatibility of local and Global Langlands correspondances. Preprint 2004.

[U06] Urban, E., Eigenvarieties for reductive groups. Preprint, 2006.

Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109-1043, U.S.A.

E-mail: cskinner@umich.edu

Department of Mathematics, Columbia University, 2990 Broadway, New York, NY 10027, U.S.A.

E-mail: urban@math.columbia.edu

Special values of L-functions modulo p

Vinayak Vatsal

Abstract. This article surveys the various known results on non-vanishing of special values of L-functions in p-adic families, with an emphasis on the rigidity theorems that underlie the proof in each case.

Mathematics Subject Classification (2000). Primary 14G10; Secondary 37A45.

Keywords. L-functions, ergodic theory.

1. Introduction

The original intent of this article was to survey the results of the author on the nonvanishing of p-adic families of anticyclotomic twists of modular L-functions of GL_2, and in particular, the introduction of Ratner's theorems in ergodic theory to this domain. However, in preparing the article, it soon became evident that the use of Ratner's theorem is an instance of an apparently general phenomenon – namely, that every result thus far known (to the author, at least) about non-vanishing modulo p of L-functions in p-adic families seems to ultimately rely on some kind of ergodic principle about the closure of certain group action orbits, of which Ratner's theorem is a sophisticated example. Since the particular subject of anticyclotomic twists has been amply described elsewhere (notably in the introduction to [7]), the present article will focus instead on surveying the general issue of nonvanishing of p-adic families of twists in variety of different settings, with the goal of exposing the common theme of rigidity which seems to underpin the whole subject. This approach may perhaps be interesting to a wider audience, and in any case may have historical legitimacy since it is the observation that orbit closures of group actions played a key role in the classical theorems of Ferrero and Washington that led the author to introduce ergodic theory in the more general setting.

1.1. Non-vanishing of twists in general. Let $\zeta(s) = \sum_{n\geq 1} n^{-s}$ denote the Riemann zeta function. This series is convergent when the real part of s is greater than 1, and admits a meromorphic continuation to $s \in \mathbb{C}$, with a simple pole at $s = 1$. Furthermore, $\zeta(s)$ satisfies the functional equation

$$\pi^{-s/2}\Gamma(s/2)\zeta(s) = \pi^{-(1-s)/2}\Gamma((1-s)/2)\zeta(1-s), \tag{1}$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

It follows trivially from this functional equation that $\zeta(k) = 0$ whenever k is a negative even integer. On the other hand, it was known to Euler that the value of $\zeta(k)$ is rational when k is negative and *odd*. For instance, the well-known formulae $\zeta(2) = \sum_n 1/n^2 = \pi^2/6$ and $\zeta(4) = \sum 1/n^4 = \pi^4/90$, together with the functional equation, show that

$$\zeta(-1) = -1/12 \quad \text{and} \quad \zeta(-3) = 1/60.$$

More generally, it can be shown that if k is a positive integer, then

$$\zeta(1-2k) = -B_k/k$$

where B_k is the Bernoulli number defined by the Taylor expansion

$$\frac{t}{e^t - 1} = \sum_{k=0}^{\infty} B_k \frac{t^k}{k!}.$$

The Bernoulli numbers B_k are closely related to the arithmetic of the cyclotomic fields $\mathbb{Q}(\zeta_p)$, where $\zeta_p = e^{2\pi i/p}$. In fact, one has the following

Theorem 1.1 (Kummer). *Let p denote an odd prime number. Then the class number of the cyclotomic field $\mathbb{Q}(\zeta_p)$ is divisible by p if and only if p divides the numerator of some B_k, for $k = 2, 4, 6, \ldots, p-3$.*

More generally, let $N > 2$ denote an integer, and let $\chi : (\mathbb{Z}/N\mathbb{Z})^\times \to \mathbb{C}^\times$ denote a primitive Dirichlet character modulo N. Then it can be shown that if $n \geq 1$ is an integer, then

$$L(1-n, \chi) = -B_{n,\chi}/n \tag{2}$$

where the twisted Bernoulli number $B_{n,\chi}$ is the algebraic number defined by the formula

$$\sum_{a=1}^{N} \frac{\chi(a) t e^{at}}{e^{Nt} - 1} = \sum B_{n,\chi} \frac{t^n}{n!}. \tag{3}$$

Furthermore, the class number formula, due to Dirichlet, relates the quantities $B_{1,\chi}$ with the class number of certain cyclotomic fields. For instance, suppose that $\chi = \chi_p$ is the quadratic residue character associated to the imaginary quadratic field $\mathbb{Q}(\sqrt{-p})$, where $p > 3$ is a prime . Then one can show that the class number $h(\mathbb{Q}(\sqrt{-p}))$ is given by

$$h(\mathbb{Q}(\sqrt{-p})) = \frac{\sqrt{p}}{\pi} \cdot L(1, \chi).$$

In this article, we will consider the general issue of determining whether some fixed prime p divides the special values of L-functions as above. In view of the class number formulae, this gives information on whether or not certain class numbers are divisible by p. For any given character, this is of course a hopeless problem, so one is naturally led to pose the following question: Suppose that S is a family of Dirichlet characters, and that n is a positive integer. Then how often is the number $L(1-n, \chi)$ divisible by a fixed prime $\mathfrak{p}$ of $\overline{\mathbb{Q}}$?

Example 1.2. Suppose that $S = S_{\text{quad}}$ is the family of quadratic characters associated to imaginary quadratic fields. Then Gauss showed that the 2-primary subgroup of the class group of $\mathbb{Q}(\sqrt{-D})$ has order 2^{g-1}, where g is the number of distinct primes dividing the discriminant of $\mathbb{Q}(\sqrt{-D})$. In particular, $h(\mathbb{Q}(\sqrt{-D}))$ is even unless D is a prime congruent to 1 mod 4.

What about other primes? If $p = 3$, Davenport and Heilbronn showed that $h(\mathbb{Q}(\sqrt{-D}))$ is prime to 3 for a positive proportion of D. For $p > 3$ it is known that there are infinitely many D with $(h(\sqrt{-D}), p) = 1$, and also infinitely many D with $(h(\sqrt{-D}), p) = p$.

Example 1.3. Recent work of Bhargava shows that at least 75% of totally real cubic fields and 50% of complex cubic fields have odd class number. For more in this direction, we refer to [2].

The examples cited above give information on the p-divisibility of various class numbers, and in view of the class number formulae, may be translated into statements about L-functions. However, it is to be noted that the *proofs* of these results are based essentially on the study of homogeneous forms of various degree, and make no reference to the L-functions as such. In the rest of this paper, we will restrict our attention to examples where one can study the L-functions directly. Specifically, we will consider the divisibility by a prime ℓ of L-functions varying in certain p-adic families. Here ℓ may or may not be the same as p.

2. p-adic families

2.1. Cyclotomic Dirichlet characters and the work of Ferrero–Washington

Example 2.1. Thus, for a different kind of example, we now take $S = S_{p\text{-cyc}}$ to denote the set of Dirichlet characters of conductor p^n, for $n \geq 0$. Such characters are in bijective correspondence with characters of the group $\text{Gal}(\mathbb{Q}(\mu_{p^\infty})/\mathbb{Q})$, where $\mathbb{Q}(\mu_{p^\infty})$ is the field obtained by adjoining to $\mathbb{Q}$ all p-power roots of unity. Thus $\mathbb{Q}(\mu_{p^\infty})$ is the union of the fields $K_n = \mathbb{Q}(\zeta_{p^n})$, where ζ_{p^n} is a primitive p^n-th root of unity. Let $h(K_n)$ denote the class number of K_n. Then one can ask how often $h(K_n)$ is divisible by a fixed prime ℓ. It turns out that the behavior depends basically on whether or not $\ell = p$.

We consider first the case that $\ell = p$. In this case, it was shown by Iwasawa that if p^{e_n} denotes the exact power of $\ell = p$ dividing the class number $h(K_n)$, then there exist integers λ, μ, and ν, such that

$$e_n = \lambda n + \mu p^n + \nu.$$

for all n sufficiently large. Iwasawa conjectured further that in fact $\mu = 0$, so that e_n is a linear function of n, which is constant if and only if $\lambda = 0$. On the other hand, experimental evidence suggests that $\text{ord}_\ell(h(K_n))$ is bounded if $\ell \neq p$. Both these phenomena were confirmed by Ferrero and Washington.

Theorem 2.2 (Ferrero–Washington). *Suppose that p is a prime number. Then the invariant μ vanishes, so we have $e_n = \lambda n + \nu$, for sufficiently large n. If $\ell \neq p$ is a fixed prime, then* $\mathrm{ord}_\ell(h(K_n))$ *is bounded as n tends towards infinity.*

We now want to make some remarks about the proof of the Ferrero–Washington theorems, since this will be the first appearance in the subject of ideas from ergodic theory.

As we have remarked above, the first step in Ferrero–Washington is to express the class numbers in terms of L-values. In view of the formula (2), the problem becomes one of determining the divisibility properties of the numbers $B_{n,\chi}$ defined in (3). In the original papers [11] and [36], the authors use an ingenious formula (apparently due to Iwasawa) which expresses the numbers $B_{n,\chi}$ in terms of the p-adic digits of certain p-adic numbers related to the $p-1$-st roots of unity. The calculation is somewhat involved, and we will not reproduce it here. But the central point may be succinctly described: to obtain the properties stated in Theorem 2.2, one needs to show that the digits of certain r-tuples of p-adic numbers behave like independent random variables.

To state this precisely, recall that $\beta \in \mathbb{Z}_p$ is called *normal* if the digits in the p-adic expansion of β contain every random string of length k with asymptotic frequency p^{-k}. It is not hard to see that β is normal in this sense if and only if the sequence of numbers $x_n(\beta) = p^{-n}s_n(\beta)$ is uniformly distributed mod 1, where $s_n(\beta)$ denotes the unique integer in the range $[0, p^n - 1]$ such that $s_n(\beta) \equiv \beta \pmod{p^n}$.

Now the main lemma in Ferrero–Washington may be stated as follows:

Lemma 2.2.1 ([11]). *Suppose that $\gamma_1, \gamma_2, \ldots, \gamma_r \in \mathbb{Z}_p$ are linearly independent over $\mathbb{Q}$. Then for almost all $\beta \in \mathbb{Z}_p$ the sequence of vectors*

$$X_n(\beta) = (x_n(\beta\gamma_1), \ldots, x_n(\beta\gamma_r)) \in [0, 1)^r$$

is uniformly distributed mod 1.

In practice, the numbers $\gamma_1, \ldots, \gamma_r$ are taken to be a maximal set of linearly independent $p-1$-st roots of unity. The connection with ergodic theory comes by analogy with the classical result of Kronecker:

Theorem 2.3 (Kronecker). *Suppose that $\gamma_1, \ldots, \gamma_r$ are real numbers, linearly independent over $\mathbb{Q}$. Then the image of the* 1*-parameter group $(t\gamma_1, \ldots, t\gamma_r)$ for $t \in \mathbb{R}$ is dense in the torus $\mathbb{R}^r/\mathbb{Z}^r$. More generally, for arbitrary γ_i, the closure of the group $(t\gamma_1, \ldots, t\gamma_r)$ is a subtorus of rank equal to the $\mathbb{Q}$-rank of the vector space spanned by the γ_i over $\mathbb{Q}$.*

Another view of the Ferrero–Washington theorems was given by Sinnott in [31] and [32], where it was observed that one can relate the Bernoulli numbers to the derivatives of certain rational functions. (This was already known to Euler.) Letting $\mathbb{F}_p$ denote the finite field with p elements, and letting $\mathbb{F}((T-1))$ denote the field of Laurent expansions in the variable $T-1$, the key lemma takes the following form:

Lemma 2.3.1 (Sinnott). *Suppose that $\gamma_1, \gamma_2, \ldots, \gamma_r \in \mathbb{Z}_p$ are linearly independent over $\mathbb{Q}$. Then the power series $T^{\gamma_1}, T^{\gamma_2}, \ldots, T^{\gamma_r}$ are algebraically independent in $\mathbb{F}_p((T-1))$.*

Here we understand that $T^a = \sum_{n=0}^{\infty} \binom{a}{n}(T-1)^n$ for any $a \in \mathbb{Z}_p$, where $\binom{a}{n} = a \cdot (a-1)\ldots(a-n+1)$.

Remark 2.4. We would like to point out here that the main ingredient in the proof of Sinnott's lemma is quite elementary and amounts to an application of Artin's theorem on the linear independence of characters. In particular, the use of explicit ergodic theory is completely absent. However, the statement that the T^{γ_i} are algebraically independent may be rephrased as stating that the ring $\mathbb{F}_p[T^{\gamma_1}, \ldots, T^{\gamma_r}] \subset \mathbb{F}_p[[T-1]]$ is isomorphic to a polynomial ring in r variables. Since $\mathbb{F}_p[[T-1]]$ is complete along the ideal $(T-1)$, and Specf$(\mathbb{F}_p[[T-1]])$ is a formal torus, this statement is formally analogous to Kronecker's theorem above in the sense that the image of the 1-parameter formal torus is Zariski dense in the r-dimensional variety Spec$(\mathbb{F}_p[T^{\gamma_1}, \ldots, T^{\gamma_r}])$.

2.2. CM L-functions. In this section we discuss the case of Hecke L-series associated to imaginary quadratic and more general CM fields. Thus let F denote a totally real field, and let M/F denote a totally imaginary quadratic extension of F. Let $\lambda\colon M^\times \backslash \boldsymbol{A}_M^\times \to \mathbb{C}$ denote an arithmetic idele class character of M. Let λ_∞ denote the restriction of λ to $(M \otimes \mathbb{R})^\times$ and write

$$\lambda_\infty(x) = \prod_\sigma \sigma(x)^{\kappa_\sigma}$$

where the product is taken over all embeddings $\sigma\colon M \to \mathbb{C}$. The formal sum $\kappa = \sum \kappa_\sigma \cdot \sigma$ is called the infinity type of λ. Let $\mathfrak{f}$ denote the conductor of λ, so that $\mathfrak{f}$ is the largest ideal of the ring of integers $\mathcal{O}_M$ with the property that $\lambda(x) = 1$ for all $x \in \mathcal{O}_M \otimes \hat{\mathbb{Z}}$ such that $x \equiv 1 \pmod{\mathfrak{f}}$.

Now let $L(s, \lambda)$ denote the L-function associated to the idele class character λ. It is well-known that the values $L(0, \lambda)$ are critical in the sense of Deligne [9], under some suitable condition on the infinity type of λ (see [8] for the case $F = \mathbb{Q}$, or [30] in general). In other words, there exists a period Ω_λ associated to λ such that the number

$$L^{\mathrm{alg}}(0, \lambda) = \frac{L(0, \lambda)}{\Omega_\lambda}$$

is an algebraic number. If we fix embeddings i_∞ and i_p of $\overline{\mathbb{Q}}$ in to $\mathbb{C}$ and $\mathbb{C}_p$ respectively, we may regard the complex number $L^{\mathrm{alg}}(0, \lambda)$ as being an element of $\mathbb{C}_p$, via the map $i_p \circ i_\infty^{-1}$. Furthermore, if one normalizes the period Ω_λ in some canonical way, one can even show that the number L^{alg} is p-adically integral in $\mathbb{C}_p$, and one can then ask whether these numbers are p-adic units, as λ varies over the members of some family. This problem was first studied by Gillard [13], [14], and Schneps [29], in the case of $F = \mathbb{Q}$ using a generalization of Sinnott's method, and the connection

between the L-values and explicit elliptic units. Further results in the case $F = \mathbb{Q}$ were given by Finis in [12]. Recently the subject was taken up by Hida in a series of deep papers (see [19], [20], for example) which treat the subject in great generality. For a sample of Hida's results, we restrict ourselves to a (relatively) simple statement. But to state even this, we need to introduce some notation. Here we follow [20].

Let X denote a finite set of embeddings $M \to \mathbb{C}$ of cardinality $[F : \mathbb{Q}]$, such that $X \cap cX = \emptyset$, where cX denotes the set $\{c\sigma,\ \sigma \in X\}$, and c denotes complex conjugation. We say that X is a p-ordinary CM-type of M if the set $\{i_p \circ i_\infty^{-1} \circ \sigma\}_{\sigma \in X}$ consists of $[F : \mathbb{Q}]$ distinct p-adic places of M. Then we consider a character λ as above whose infinity type is given by

$$k \sum_{\sigma \in X} \sigma + \kappa(1 - c)$$

where $\kappa = \sum_{\sigma \in X} \kappa_\sigma \sigma$ with $\kappa_\sigma \geq 0$ for $\sigma \in X$, and $0 < k \in \mathbb{Z}$. Then to define the transcendental factor Ω_λ, we can proceed as follows. Pick an abelian variety A of CM type such that $A(\mathbb{C}) \cong \mathbb{C}^{[F:\mathbb{Q}]}/\mathfrak{a}$, where the product is indexed by the $[F : \mathbb{Q}]$ places in X, and the fractional ideal $\mathfrak{a} \subset \mathcal{O}_M$ is embedded diagonally via the corresponding places of X. Let $R \subset \overline{\mathbb{Q}}$ denote the Witt ring of $\mathcal{O}_M$, with respect to the place induced by i_p. Then A can be defined over R, and we can pick a Néron differential ω on A such that ω generates $\Omega_{A/W}$ over R. Picking an isomorphism $\phi\colon A(\mathbb{C}) \cong \mathbb{C}^{[F:\mathbb{Q}]}/\mathfrak{a}$, we define a vector $\Omega_\infty \in \mathbb{C}^{[F:\mathbb{Q}]}$ via $\phi^*\big(\prod(du_\sigma)\big) = \Omega_\infty \omega$. Here u_σ is the standard complex variable on the copy of $\mathbb{C}$ indexed by σ. Writing the components of Ω_∞ as Ω_σ for $\sigma \in X$, we have that $\Omega_\sigma \neq 0$, and

$$L^{\mathrm{alg}}(0, \lambda) = \frac{\prod_{\sigma \in X} \pi^{\kappa_\sigma} \Gamma(k + \kappa_\sigma) L(0, \lambda)}{\prod_{\sigma \in X} \Omega_\sigma^{k+2\kappa_\sigma}} \in R \tag{4}$$

Then the problem is to study p-divisibility properties of the numbers $L^{\mathrm{alg}}(0, \lambda)$ as defined above, as λ varies over some prescribed set. Recall therefore that the character $\chi\colon \boldsymbol{A}_M^\times \to \mathbb{C}^\times$ is called anticyclotomic if $\chi \circ \tau = \chi^{-1}$, where τ is the nontrivial automorphism of M/F. Then we let λ denote a fixed Hecke character, and consider the values $L^{\mathrm{alg}}(0, \lambda\chi)$, as χ varies over the set of anticyclotomic characters of M of conductor $\mathfrak{l}^n$, for some fixed prime $\mathfrak{l}$ of K with residue characteristic $\ell \neq p$, and an integer n. The principal result in this direction is due to Hida. To state the theorem, let us write $L_{\mathfrak{l}}^{\mathrm{alg}}(0, \lambda\chi) = L^{\mathrm{alg}}(0, \lambda\chi) \cdot (1 - \lambda\chi(\mathfrak{l}))$ for the algebraic part of the $\mathfrak{l}$-imprimitive L-function.

Theorem 2.5 ([20], Theorem 1.1). *Suppose that $p > 2$ is an unramified prime in $M/\mathbb{Q}$ and that (M, X) is a p-ordinary CM type. Fix a character λ of conductor 1 with infinity type $k \sum_{\sigma \in X} \sigma + \kappa(1 - c)$ as above. Then we have*

$$|L^{\mathrm{alg}}(0, \lambda\chi)|_p = 1$$

for almost all anticyclotomic characters of $\mathfrak{l}$-power conductor, unless the following three conditions are satisfied simultaneously:

1. *M/F is unramified at all finite places,*

2. *the Artin symbol $\left(\frac{M/F}{\mathfrak{c}}\right)$ has the value -1, for the polarization ideal $\mathfrak{c}$ of A, and*

3. *for all ideals $\mathfrak{a}$ prime to p, we have $\lambda N(\mathfrak{a}) \equiv \left(\frac{M/F}{\mathfrak{c}}\right) \pmod{\mathfrak{m}}$.*

If all these conditions are satisfied simultaneously, then $|L^{\mathrm{alg}}(0, \lambda\chi)|_p < 1$ for all χ.

Here we understand that 'almost all' means 'a Zariski dense subset' in general, and 'all but finitely many' if $F_{\mathfrak{l}}$ has dimension 1 over $\mathbb{Q}_p$.

The proof of this theorem is long and intricate, and we will not discuss it here, except to remark that the proof is a generalization of Sinnott's method mentioned above. Essentially, one has to relate the values of the L-function to the values of certain Hilbert modular Eisenstein series at CM points, and prove a basic result on the linear independence of certain of these series in characteristic p. For our purposes, it will suffice to observe that key ingredient is a rigidity theorem of C.-L. Chai, which enables one to prove that under some condition, schemes fixed by torus actions tend to be rather big. For comparison with the results from ergodic theory that were cited above, we state a precise theorem, as follows.

Suppose that k is an algebraically closed field of characteristic $p > 0$ and let X be a finite dimensional smooth formal p-divisible group over k. Let $E_{\mathbb{Z}_p} = \mathrm{End}(X)$ and set $E = E_{\mathbb{Z}_p} \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$. Then E is a finite dimensional vector space over $\mathbb{Q}_p$. We let $\boldsymbol{E}$ denote the linear algebraic group over $\mathbb{Q}_p$ such that $\boldsymbol{E}(R) = (E \otimes_{\mathbb{Q}_p} R)^{\times}$ for any commutative $\mathbb{Q}_p$-algebra R.

If G is any connected algebraic group over $\mathbb{Q}_p$, and $\rho\colon G \to \boldsymbol{E}$ is a homomorphism of algebraic groups, then we may regard ρ as a linear representation of G on the vector space E via the canonical map $\boldsymbol{E} \subset \mathrm{Aut}(E)$. Then Chai has proven the following striking result. (The notation is as above.)

Theorem 2.6 (Chai). *Suppose that the trivial representation is not a subquotient of the representation ρ of G on E. Suppose also that Z is a reduced and irreducible closed formal subscheme of X which is closed under the action of an open subgroup of $G(\mathbb{Z}_p)$. Then Z is closed under the group law of X and is a p-divisible subgroup scheme of X.*

2.3. Anticyclotomic L-functions. Finally, we treat the applications of ergodic theory to anticyclotomic L-functions associated to Hilbert modular forms over totally real fields.

To describe the results, let F denote a totally real field, and let K/F denote an imaginary quadratic extension. Let π denote a cuspidal automorphic representation of $\mathrm{GL}_2(F)$. We assume throughout that the data of π and K are *non-exceptional*, meaning that the representations π and $\pi \otimes \eta$ are distinct, where η denotes the quadratic character associated to the extension K/F.

Let $\chi : \boldsymbol{A}_K^\times / K^\times \to \mathbb{C}$ be a quasi-character of K, and write $L(\pi, \chi, s)$ for the Rankin–Selberg L-function associated to π and $\pi(\chi)$. Here $\pi(\chi)$ denotes the automorphic representation of GL_2 attached to χ. (For the definitions, we refer the reader to [22] and [21].) Then this L-function, which is first defined as a product of Euler factors over all places of F, may be shown to have a meromorphic extension to $\mathbb{C}$ with functional equation

$$L(\pi, \chi, s) = \varepsilon(\pi, \chi, s) L(\tilde{\pi}, \chi^{-1}, 1 - s)$$

where $\tilde{\pi}$ is the contragredient of π and $\varepsilon(\pi, \chi, s)$ is the ε-factor.

Let $\omega : \boldsymbol{A}_F^\times / F^\times \to \mathbb{C}^\times$ be the central quasi-character of π. We will make the following assumption on the quasi-characters ω and χ:

$$\chi \cdot \omega = 1 \quad \text{on } \boldsymbol{A}_F^\times \subset \boldsymbol{A}_K^\times. \tag{5}$$

This assumption implies that $L(\pi, \chi, s)$ is entire and equal to $L(\tilde{\pi}, \chi^{-1}, s)$. Thus the functional equation of $L(\pi, \chi, s)$ may be restated as

$$L(\pi, \chi, s) = \varepsilon(\pi, \chi, s) L(\pi, \chi, 1 - s)$$

and the parity of the order of vanishing of $L(\pi, \chi, s)$ at $s = 1/2$ is determined by the value of

$$\varepsilon(\pi, \chi) \stackrel{\text{def}}{=} \varepsilon(\pi, \chi, 1/2) \in \{\pm 1\}.$$

Following [6] and [7], we say that the pair (π, χ) is *even* or *odd*, depending upon whether $\varepsilon(\pi, \chi)$ is $+1$ or -1. According to the conjectures introduced by Mazur in[24], it is expected that the order of vanishing of $L(\pi, \chi, s)$ at $s = 1/2$ should 'usually' be minimal, meaning that either $L(\pi, \chi, 1/2)$ or $L'(\pi, \chi, 1/2)$ should be nonzero, depending upon whether (π, χ) is even or odd.

Results of this kind were first proven by Rohrlich [28], for the case where $F = \mathbb{Q}$, and π and K are *exceptional* in the sense that $\pi \cong \pi \otimes \eta$, using results from transcendence theory, notably p-adic cases of Roth's theorems. However, nothing was known for non-exceptional π and K until the introduction of ergodic theory in [33] and [34], which treated the case of $F = \mathbb{Q}$. The ideas from ergodic theory were quickly assimilated and extended in [5], and the generalization to the case of general F was given in [6] and [7].

To proceed, we need to introduce some notation. Thus let $\mathfrak{n}$ denote the conductor of the representation π. Let $\mathfrak{p}$ denote a fixed prime of F, and let χ denote a ring class character of $\mathfrak{p}$-power conductor. Here we recall that the quasi-character χ of K is called a ring class character, or an anticyclotomic character, if the restriction of χ to $\boldsymbol{A}_F^\times$ is everywhere unramified. Then we propose to study the order of vanishing of $L(\pi, \chi, s)$ at $s = 1/2$ as χ varies over the set $S = S_{\mathfrak{p}}^{\text{anticyc}}$ of ring class characters of $\mathfrak{p}$-power conductor.

In view of equation (5), it makes sense to require that the central character ω of π is everywhere unramified. We assume also that π corresponds to a Hilbert modular

form of parallel weight $(2, \ldots, 2)$ and that the discriminant $\mathfrak{D}$ of K/F is relatively prime to the prime-to-$\mathfrak{p}$ part $\mathfrak{n}'$ of $\mathfrak{n}$. Under these conditions, it may be shown that for all $n \gg 0$, and all χ of conductor $\mathfrak{p}^n$, the root number $\varepsilon(\pi, \chi)$ is given by the formula

$$\varepsilon(\pi, \chi) = (-1)^{\#S} \tag{6}$$

where S denotes the set of real places of F, together with those finite primes of F which do not divide $\mathfrak{p}$, are inert in K, and divide $\mathfrak{n}$ to an odd power. In particular, the root number $\varepsilon = \varepsilon(\pi, \chi) = \pm 1$ depends only on π and K and $\mathfrak{p}$, once the conductor of χ is sufficiently divisible. Thus one expects the order of vanishing of $L(\pi, \chi, 1/2)$ to be equal to either 0 or 1, for 'generic' χ, depending only on the sign of ε. That this is indeed the case was confirmed by the main results in [6] and [7], and we refer the reader to the introduction of [6] for a very detailed discussion.

In the present paper, we will focus on the non-vanishing of $L(\pi, \chi, 1/2)$ modulo a prime of $\overline{\mathbb{Q}}$. The basic results in this direction were given in [34], for the case $F = \mathbb{Q}$, and we now proceed to state them.

Thus, let us assume that $F = \mathbb{Q}$. Let $N = \mathfrak{n}$ denote the level of π, let $D = \mathfrak{D}$ denote the discriminant of the imaginary quadratic field $K = \mathbb{Q}(\sqrt{D})$, and let $p = \mathfrak{p}$ denote a rational prime. We assume further that the numbers N, D, p are pairwise co-prime. We let f denote the primitive form of level N associated to π; since the central character ω of π is unramified and $\mathbb{Q}$ has class number 1, we see that f is a primitive form on the group $\Gamma_0(N)$. We assume further that we are in the even case, so that there are an even number of places in S. We let $\Omega_\pi = \Omega_f$ denote the canonical integral period for f, as defined by Hida in [18]. Then the number

$$L^{\mathrm{alg}}(\pi, \chi) = \frac{L(\pi, \chi, 1/2)}{\Omega_\pi} \cdot C_\chi$$

is an algebraic integer. Here $C_\chi = Dp^{2n}$, where p^n denotes the conductor of χ. Let λ denote a fixed prime of $\overline{\mathbb{Q}}$, and consider the λ-adic absolute value $|L^{\mathrm{alg}}(\pi, \chi)|_\lambda$. We want to study the general question of how $|L^{\mathrm{alg}}(\pi, \chi)|_\lambda$ varies as a function of χ, and the result depends on whether or not λ has residue characteristic p. In either case, let us define two constants C_{csp} and C_{Eis} associated to π, as in [34], Section 2.4.[1]

Then one has the following result:

Theorem 2.7 ([34]). *Suppose that λ has residue characteristic $\ell \neq p$. Then we have*

$$|L^{\mathrm{alg}}(\pi, \chi)|_\lambda = |C_{\mathrm{csp}}^2 C_{\mathrm{Eis}}|_\lambda$$

for all but finitely many λ of conductor p^n.

Actually, the theorem above was stated in [34] under some mild assumptions on ℓ, but these restrictions are easily removed, for example with the improved formalism

[1] The definition of these constants is rather technical, and we prefer not to reproduce it. The significance of these numbers, in particular the relationship to congruences, is elucidated in the paper [25].

introduced in [7], or by a slightly more detailed analysis of the original proof. We remark here that the numbers C_{csp} and C_{Eis} are *not* necessarily λ-adic units.

As for the case where λ has residue characteristic p, the result is in the same vein, provided one assumes that the local component π_p is ordinary at λ, in the sense that the Hecke eigenvalue $a_p(\pi)$ is a λ-adic unit.

Theorem 2.8 ([34]). *Suppose that λ has residue characteristic $\ell = p$ and that π_p is ordinary at λ. Then we have*

$$\lim_{\chi} |L^{\mathrm{alg}}(\pi, \chi)|_\lambda = |C_{\mathrm{csp}}^2 C_{\mathrm{Eis}}|_\lambda,$$

where the limit is taken over characters χ of conductor p^n, as $n \to \infty$.

Remark 2.9. In view of recent results in the Iwasawa theory of elliptic curves, our results on L-functions may be formulated in terms of the growths of certain Selmer groups, which are generalizations of the Iwasawa ideal class groups occurring in our discussion of the Ferrero–Washington theorem above. For more details, we refer the reader to [1] and [25]. We remark also that our results above have not yet been extended to general F, but it seems likely that such generalizations would follow without difficulty from the techniques of [7].

Remark 2.10. We point out also that there are results analogous to those above in the case that the sign in the functional equation is -1. However, in these cases, one is dealing with derivatives of L-functions, and there is no general notion of what it means for a derivative of an L-function to be nonzero modulo p. In the case at hand one has an *ad hoc* definition in terms of p-divisibility of certain Heegner points arising from the Gross–Zagier formula for derivatives, and it is this kind of result that is proven. For details we refer the reader to [34] and [5].

In keeping with the general theme of this article, we wish now to elaborate on the role of ergodic theorems in the proofs of our results. A detailed description of the strategy may be found in the introductions to [33] and [7], and we will not cut and paste from those articles here. For the present, we simply note that the starting point comes from the formulae of Gross, Zagier, and Zhang, which relate the values of the L-functions in question to the heights of certain special points on quaternion algebras. (See [17], [15], and [37] for the theorems, which were then reframed in the article [16]. A more elementary perspective may be found in [35].) In view of these special value formulae, the essential point in proving that the L-values are non-zero modulo p boils down to showing that certain vectors whose components are formed by the special points and their conjugates, are uniformly distributed in the appropriate sense on certain Shimura varieties. The necessary uniform distribution is then deduced by applying deep theorems in ergodic theory due to M. Ratner [27].[2]

[2] In [6] and [7], appeals were made to results of Margulis and Tomanov [23], since these results were formulated in a manner convenient for our applications there. The author has since been informed by Ratner that the results we quoted from [23] can in fact be deduced from those given earlier in [26].

The reduction of our number theoretic results to Ratner's theorem has been amply documented elsewhere, so we will just state some of Ratner's key results, in a manner that we hope will make clear the analogy with the results of Ferrero–Washington, Sinnott, and Chai. It is perhaps germane to remark here that the introduction of Ratner's theorem in [33] was inspired by direct analogy with the use of Kronecker's theorem by Ferrero and Washington.

Thus let G denote the p-adic Lie group $\mathrm{SL}_2(\mathbb{Q}_p)$, and let $\Gamma_i \subset G$, $i = 1, 2$ denote discrete and cocompact subgroups. We say that Γ_1 and Γ_2 are commensurable if $\Gamma_1 \cap \Gamma_2$ has finite index in Γ_1 and Γ_2. Then the following remarkable result is a very simple consequence of the main results in [27]:

Theorem 2.11 (Ratner). *The set* $\Gamma_1 \cdot \Gamma_2 = \{\gamma_1 \cdot \gamma_2,\ \gamma_1 \in \Gamma_1,\ \gamma_2 \in \Gamma_2\}$ *is dense in* G *if and only if the subgroups* Γ_1 *and* Γ_2 *are not commensurable.*

Remark 2.12. Note that it is obvious that the product $\Gamma_1 \cdot \Gamma_2$ cannot be dense if Γ_1 and Γ_2 are commensurable. But the reverse implication is extremely deep, and seems to admit no elementary proof.

Remark 2.13. For an analogy, let G denote the additive group $\mathbb{R}$ of real numbers, and let X_i, $i = 1, 2$ denote discrete subgroups of G. Then each X_i is abstractly isomorphic to the additive group of $\mathbb{Z}$. If x_i is a generator of X_i, then the groups X_i are commensurable if and only if the x_i are linearly dependent over $\mathbb{Q}$. In this case the product $X_1 \cdot X_2$ is discrete in G. On the other hand, Kronecker's theorem implies that the product $X_1 \cdot X_2$ is dense if the x_i are independent over $\mathbb{Q}$, which is to say, if the groups X_i fail to be commensurable. Thus Ratner's theorem above is a p-adic and non-abelian analogue of Kronecker's theorem.

Actually, one requires a slightly more refined theorem for the applications to number theory. As above, write G for the p-adic Lie group $\mathrm{SL}_2(\mathbb{Q}_p)$. Let r denote a positive integer, and for each i with $1 \leq i \leq r$, we let Γ_i denote a discrete and cocompact subgroup of G. Then

$$\Gamma = \prod_{i=1}^{r} \Gamma_i \subset \prod_{i=1}^{r} G$$

is a discrete and cocompact subgroup of the product G^r of r copies of G. We may then formulate the following result:

Theorem 2.14 (Ratner). *Suppose that the groups* Γ_i *are pairwise non-commensurable. Then the image of the diagonal* $\Delta(G) = \{(g, \ldots, g),\ g \in G\} \subset G^r$ *has dense image in the quotient* $\Gamma \backslash G^r$.

Finally, we give a rigidity result which implies the two above as special cases.

Theorem 2.15 (Ratner). *Let* G *denote a* p*-adic Lie group, and let* $\Gamma \subset G$ *be such that* $\Gamma \backslash G$ *has finite volume with respect to the unique* G*-invariant measure. Let* $U \subset G$

denote any subgroup generated by 1*-parameter subgroups, namely, by the image of* (*additive*) *homomorphisms* $u_i : \mathbb{Q}_p \to G$. *Then the closure* $\overline{U}$ *of the orbit of* U *in* $\Gamma\backslash G$ *is homogeneous, in the sense that there exists a subgroup* H *of* G *such that the orbit of* H *is closed in* $\Gamma\backslash G$ *and* $\overline{U}$ *coincides with the orbit of* H.

Remark 2.16. We would like to point out here that for the purposes of the results in [5] and some of the results in [7] (namely, the indefinite case), one can get by with yet another kind of rigidity principle, namely certain cases of the André–Oort conjecture. For more on this subject, we refer the reader to [5]. For the relationship with Chai's theorem, see [3]. Further discussion of this and related topics may be found in [10] and [4].

To conclude, we would hope that the analogy between the theorems of Kronecker, Sinnott, Chai, and Ratner is now evident. Namely, in every case, we are asserting that the closure of rather small group orbits, (the diagonal, in Ratner's case, or a 1-parameter group in Kronecker's theorem) is forced, by rigidity, to be rather big. In the Ferrero–Washington and anticyclotomic cases, the orbit of a small group inside an r-dimensional object turns out to be dense, and in every case, including the theorem of Chai, the key statement is a rigidity principle of the form that the closures of the relevant orbits coincide with the orbits of subgroups. Is there a general ergodic or rigidity principle that accounts for all of these results? We hope that the answer is affirmative, but at present we seem to be far from finding it.

References

[1] Bertolini, Massimo, and Darmon, Henri, Iwasawa's main conjecture for elliptic curves over anticyclotomic extensions. *Ann. of Math.*, to appear.

[2] Bhargava, Manjul, The density of discriminants of quartic rings and fields. *Ann. of Math.* (2) **162** (2) (2005), 1031–1063.

[3] Chai, C.-L., Families of ordinary abelian varieties: canonical coordinates, p-adic monodromy, tate-linear subvarieties and hecke orbits. Preprint, 2003.

[4] Clozel, Laurent, Oh, Hee, and Ullmo, Emmanuel, Hecke operators and equidistribution of Hecke points. *Invent. Math.* **144** (2) (2001), 327–351.

[5] Cornut, C., Mazur's conjecture on higher Heegner points. *Invent. Math.* **148** (3) (2002), 495–523.

[6] Cornut, C., and Vatsal, V., Nontriviality of Rankin-Selberg L-functions and CM points. Preprint, 2004.

[7] Cornut, C., and Vatsal, V., CM points and quaternion algebras. *Doc. Math.* **10** (2005), 263–309 (electronic).

[8] de Shalit, E., *Iwasawa theory of elliptic curves with complex multiplication*. Perspect. Math. 3, Academic Press, Boston, MA, 1987.

[9] Deligne, P., Valeurs de fonctions L et périodes d'intégrales. In *Automorphic forms, representations and L-functions*, Proc. Sympos. Pure Math. 33, Part 2, Amer. Math. Soc., Providence, RI, 1979, 313–346.

[10] Edixhoven, B., and Yafaev, A., Subvarieties of Shimura varieties. *Ann. of Math.* (2) **157** (2) (2003), 621–645.

[11] Ferrero, B., and Washington, L., The Iwasawa invariant μ_p vanishes for abelian number fields. *Ann. of Math.* (2) **109** (2) (1979), 377–395.

[12] Finis, T., Divisibility of anticyclotomic L-functions and theta functions with complex multiplication. Preprint, 2002.

[13] Gillard, Roland, Fonctions L p-adiques des corps quadratiques imaginaires et de leurs extensions abéliennes. *J. Reine Angew. Math.* **358** (1985), 76–91.

[14] Gillard, Roland, Croissance du nombre de classes dans des $\mathbf{Z}_l$-extensions liées aux corps quadratiques imaginaires. *Math. Ann.* **279** (3) (1988), 349–372.

[15] Gross, B., Heights and the special values of L-series. In *Number Theory* (ed. by H. Kisilevsky and J. Labute,), CMS Conference Proceedings 7, Amer. Math. Soc., Providence, RI, 1987, 115–189.

[16] Gross, B., Heegner points and representation theory. In *Heegner points and Rankin L-series*, Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 37–66.

[17] Gross, B., and Zagier, D., Heegner points and derivatives of L-series. *Invent. Math.* **84** (1986), 225–320.

[18] Hida, H., Modules of congruence of Hecke algebras and L-functions associated with cusp forms. *Amer. J. Math.* **110** (2) (1988), 323–382.

[19] Hida, H., The Iwasawa μ-invariant of p-adic Hecke L-functions. Preprint, 2006.

[20] Hida, H., Non-vanishing modulo p of Hecke L-values. In *Geometric Aspects of Dwork Theory* (ed. by A. Adolphson, F. Baldassarri, P. Berthelot, N. Katz and F. Loeser), Volume II, Walter de Gruyter, Berlin 2004, 731–780.

[21] Jacquet, H., *Automorphic forms on* GL(2). *Part II*, Lecture Notes in Math. 278, Springer-Verlag, Berlin 1972.

[22] Jacquet, H., and Langlands, R. P., *Automorphic forms on* GL(2). Lecture Notes in Math. 114, Springer-Verlag, Berlin 1970.

[23] Margulis, G. A., and Tomanov, G. M., Invariant measures for actions of unipotent groups over local fields on homogeneous spaces. *Invent. Math.* 116(1-3) (1994), 347–392.

[24] Mazur, B., Modular curves and arithmetic. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 185–211.

[25] Pollack, R., and Weston, T., On mu-invariants of anticyclotomic p-adic L-functions of elliptic curves. Preprint, 2005.

[26] Ratner, M., Raghunathan's Conjectures for p-adic Lie groups. *Internat Math. Res Notices* **5** (1993), 141–146.

[27] Ratner, M., Raghunathan's conjectures for Cartesian products of real and p-adic Lie groups. *Duke Math. J.* **77** (2) (1995), 275–382.

[28] Rohrlich, D. E., On L-functions of elliptic curves and cyclotomic towers. *Invent. Math.* **75** (3) (1984), 409–423.

[29] Schneps, Leila, On the μ-invariant of p-adic L-functions attached to elliptic curves with complex multiplication. *J. Number Theory* **25** (1) (1987), 20–33.

[30] Shimura, Goro, On some arithmetic properties of modular forms of one and several variables. *Ann. of Math.* (2) **102** (3) (1975), 491–515.

[31] Sinnott, W., On the μ-invariant of the Γ-transform of a rational function. *Invent. Math.* **75** (2) (1984), 273–282.

[32] Sinnott, W., Γ-transforms of rational function measures on $\mathbf{Z}_S$. *Invent. Math.* **89** (1) (1987), 139–157.

[33] Vatsal, V., Uniform distribution of Heegner points. *Invent. Math.* **148** (2002), 1–46.

[34] Vatsal, V., Special values of anticylotomic L-functions. *Duke Math J.* **116** (2) (2003), 219–261.

[35] Vatsal, Vinayak, Special value formulae for Rankin L-functions. In *Heegner points and Rankin L-series*, Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 165–190.

[36] Washington, L., The non-p-part of the class number in a cyclotomic $\mathbb{Z}_p$-extension. *Invent. Math.* **49** (1) (1978), 87–97.

[37] S. Zhang, S., Gross-Zagier formula for GL_2. II. In *Heegner points and Rankin L-series*, Math. Sci. Res. Inst. Publ. 49, Cambridge University Press, Cambridge 2004, 191–242.

Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada

E-mail: vatsal@math.ubc.ca

Higher-dimensional analogues of stable curves

Valery Alexeev

Abstract. The Minimal Model Program offers natural higher-dimensional analogues of stable n-pointed curves and maps: stable pairs consisting of a projective variety X of dimension ≥ 2 and a divisor B, that should satisfy a few simple conditions, and stable maps $f\colon (X, B) \to Y$. Although MMP remains conjectural in higher dimensions, in several important situations the moduli spaces of stable pairs, generalizing those of Deligne–Mumford, Knudsen and Kontsevich, can be constructed more directly, and in considerable generality. We review these constructions, with particular attention paid to varieties with group action, and list some open problems.

Mathematics Subject Classification (2000). Primary 14J10; Secondary 14E30, 14L30.

Keywords. Moduli spaces, stable pairs, Minimal Model Program.

Introduction

Stable curves were introduced by Deligne and Mumford in [19] and proved to be extremely useful, with diverse applications in many fields of mathematics and in physics. Stable maps from n-pointed curves to varieties were used by Kontsevich to define Gromov–Witten invariants. The study of the moduli spaces of stable curves and maps is a thriving field.

Stable surfaces, the two-dimensional analogues of stable curves, were introduced by Kollár and Shepherd-Barron in [39]. It was consequently realized [4], [3] that this definition can be extended to higher-dimensional varieties and, moreover, to pairs (X, B), consisting of a projective variety X of dimension ≥ 2 and a divisor B, and to stable maps $f\colon (X, B) \to Y$. One arrives at this definition by mimicking the construction of stable one-parameter limits of curves in the higher-dimensional case, and replacing contractions of (-1)- and (-2)-curves by the methods of the Minimal Model Program.

Stable pairs provide an apparently very general, nearly universal way to compactify moduli spaces of smooth or mildly singular varieties and pairs. There are, however, two complications. First, as of this writing, the Minimal Model Program in arbitrary dimensions is still conjectural. Secondly, even in the case of surfaces the resulting moduli spaces turn out to be very complicated, and numerical computations similar to the curve case seem to be out of reach.

The situation can be improved in both respects by looking at some particularly nice classes of varieties, such as abelian varieties and other varieties with group action:

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

toric, spherical, and also at varieties and pairs closely related to them, for example the hyperplane arrangements.

In all of these cases the Minimal Model Program can be used for guessing the correct answer, but the actual constructions of the moduli spaces can be made without it, by exploiting symmetries of the situation. At the same time, the resulting moduli spaces come equipped with rich combinatorial structures, typically with stratifications labeled by various polytopal tilings.

The aim of this paper is to review the basic constructions and several of the examples mentioned above. My understanding of the subject was shaped over the years by discussions with (in the chronological order) J. Kollár, S. Mori, I. Nakamura, K. Hulek, Ch. Birkenhake, M. Brion, B. Hassett, A. Knutson, and many other people whom I am unable to list here. I am indebted to them all.

1. Definition of stable pairs and maps

To define varieties and pairs, we work over an algebraically closed field k of arbitrary characteristic. All *varieties* will be assumed to be connected and reduced but not necessarily irreducible. A *polarized variety* is a projective variety X with an ample invertible sheaf L.

Definition 1.1. Let X be a projective variety, B_j, $i = 1, \dots, n$, be effective Weil divisors on X, possibly reducible, and b_j be some rational numbers with $0 < b_j \leq 1$. The pair $(X, B = \sum b_j B_j)$ (resp. a map $f\colon (X, B) \to Y$) is called *stable* if the following two conditions are satisfied:

1. *on singularities:* the pair (X, B) is semi log canonical, and
2. *numerical:* the divisor $K_X + B$ is ample (resp. f-ample).

Both parts require an explanation.

Definition 1.2. Assume that X is a *normal* variety. Then X has a canonical Weil divisor K_X defined up to linear equivalence. The pair (X, B) is called *log canonical* if

1. $K_X + B$ is $\mathbb{Q}$-Cartier, i.e. some positive multiple is a Cartier divisor, and
2. for every proper birational morphism $\pi\colon X' \to X$ with normal X', in the natural formula

$$K_{X'} + \pi_*^{-1} B = \pi^*(K_X + B) + \sum a_i E_i$$

one has $a_i \geq -1$. Here, E_i are the irreducible exceptional divisors of π, and the pullback π^* is defined by extending $\mathbb{Q}$-linearly the pullback on Cartier divisors. $\pi_*^{-1} B$ is the strict preimage of B.

If $\operatorname{char} k = 0$ then X has a resolution of singularities $\pi\colon X' \to X$ such that $\operatorname{Supp}(\pi_*^{-1} B) \cup E_i$ is a normal crossing divisor; then it is sufficient to check the condition $a_i \geq -1$ for this morphism π only.

The definition for semi log canonical surface singularities X originated in [39]. The following definition, equivalent to [39] in the surface case, and extending it to higher-dimensional varieties and pairs, is from [3].

Definition 1.3. A pair (X, B) is called *semi log canonical* if

1. X satisfies Serre's condition S2, in particular, equidimensional,
2. X has at worst double normal crossing singularities in codimension one, and no divisor B_j contains any component of this double locus,
3. some multiple of the Weil $\mathbb{Q}$-divisor $K_X + B$, well defined thanks to the previous condition, is $\mathbb{Q}$-Cartier, and
4. denoting by $\nu : X^\nu \to X$ the normalization, the pair $(X^\nu,\ (\text{double locus}) + \nu^{-1}B)$ is log canonical.

Example 1.4. Assume that X is a curve. Then (X, B) is semi log canonical iff X is at worst nodal, B_j do not contain any nodes, and for every $P \in X$ one has $\operatorname{mult}_P B = \sum b_j \operatorname{mult}_P B_j \leq 1$. A map $f\colon (X, B) \to Y$ is stable if, in addition to this condition on singularities, the divisor $K_X + B$ has positive degree on every irreducible component of X collapsed by f.

Hence, for $b_j = 1$, $\deg B_j = 1$, and $Y =$ a point (i.e. in the absolute case) these are precisely the Deligne–Mumford–Knudsen stable n-pointed curves [19], [33], [32]. With the same assumptions on B but Y arbitrary, these are Kontsevich's stable maps. Hassett [27] considered the absolute case with $0 < b \leq 1$, $\deg B_j = 1$, for which he constructed a smooth Deligne–Mumford stack with a projective moduli space.

The motivation for the definition of stable pairs is that they appear as natural limits of one-parameter families of smooth varieties and pairs $(X, B) \to S$, as will be developed in Section 2. In higher dimensions, there is an additional complication: if the total divisor B is not $\mathbb{Q}$-Cartier then the central fiber B_0 may have an embedded component, so no longer be an ordinary divisor. There are several ways to fix this:

1. Pairs with floating coefficients b_j. We will say that a pair
$$\Bigl(X,\ B = \sum b_j B_j + \sum (b_k + \varepsilon_k) B_k\Bigr)$$
(resp. a map) is stable if, in addition, the divisors B_k in the second group are $\mathbb{Q}$-Cartier and for all $0 < \varepsilon_k \ll 1$, the pair $\bigl(X, \sum b_j B_j + \sum (b_k + \varepsilon_k) B_k\bigr)$ is stable.
2. Pairs with coefficients b_j outside of a "bad" subset of $[0, 1]$. Again, the idea here is the same as in (1), to avoid the values b_j for which the total divisors B_j may be not $\mathbb{Q}$-Cartier.
3. Working with subschemes $B_j \subset X$ instead of simply divisors.
4. Working with finite morphisms $B_j \to X$, where B_j are (reduced) varieties of dimension $\dim B_j = \dim X - 1$, rather than with embedded divisors.

2. Minimal Model Program construction

The true motivation for the introduction of stable pairs is that they inevitably appear as limits of one-parameter families of smooth varieties and pairs, when one tries to follow the classical construction in the case of curves. This is explained by the following statement, which however is conditional: it depends on the validity of the log Minimal Model Program in dimension $\dim X + 1$ (so, currently problematic for pairs (X, B) with $\dim X \geq 3$) and on Inversion of Adjunction in an appropriate sense, as explained in the sketch of the proof below. The argument also requires $\operatorname{char} k = 0$ or X to be a curve for the resolution of singularities and semistable reduction.

This statement appeared in [39] in the case of surfaces with $B = 0$, where it is not conjectural and in [3] in the more general case (see also [26]).

By a one-parameter family of stable maps we will understand a morphism $f\colon (X, B) \to Y \times S$, where $(S, 0)$ is a germ of a nonsingular curve, such that $\pi = p_2 \circ f\colon X \to S$ and $\pi|_{B_j}\colon B_j \to S$ are flat, and every geometric fiber $f_{\bar{s}}\colon (X_{\bar{s}}, B_{\bar{s}}) \to Y$ is a stable map. We will denote $S \setminus 0$ by U.

The definition of a family over an arbitrary scheme S is similar but requires care, especially if S is not reduced. We will discuss it in the next section.

Theorem 2.1 (Properness of the functor of stable maps). *Every punctured family $f_U\colon (X_U, B_U) \to Y \times U$, of stable pairs has at most one extension to a family of stable pairs over S. Moreover, such an extension does exist after a finite base change $(S', 0) \to (S, 0)$.*

Sketch of the proof. We assume that fibers X_s for $s \neq 0$ are normal, for simplicity. Denote an extension by $f\colon (X, B) \to Y \times S$. Inversion of Adjunction of Shokurov–Kollár (see, e.g. [35], 17.3) says that the central fiber (X_0, B_0) is semi log canonical iff the pair $(X, B + X_0)$ is log canonical.

Now suppose that we have an extension. Then $(X, B + X_0)$ has log canonical singularities and $K_X + B + X_0$ is f-ample. Hence, $(X, B + X_0)$ is *the log canonical model* of $(\widetilde{X}, \widetilde{B} + \widetilde{X}_{0,\mathrm{red}})$ over $Y \times S$ for any resolution of singularities $(\widetilde{X}, \widetilde{B})$ of any extension of f_U. Existence of the log canonical model is the main result of the log Minimal Model Program, and its uniqueness is a basic and easy fact, see, e.g. [38].

In the opposite direction, pick some extension family. Take a resolution of singularities, which introduces some exceptional divisors E_i. Apply the Semistable Reduction Theorem to this resolution. The result is that after a ramified base change $(S', 0) \to (S, 0)$ we now have an extended family $\tilde{f}'\colon (\widetilde{X}', \widetilde{B}')$ such that $\widetilde{X}'$ is smooth, the central fiber $\widetilde{X}'_0$ is a reduced normal crossing divisor, and, moreover, $\widetilde{X}'_0 \cup \operatorname{Supp} \widetilde{B}' \cup \widetilde{E}'_i$ is a normal crossing divisor.

It follows that the pair $(\widetilde{X}', \widetilde{B}' + \widetilde{X}'_0 + \sum \widetilde{E}'_i)$ has log canonical singularities and is relatively of general type over $Y \times S'$. Now let $f'\colon (X', B' + X'_0) \to Y \times S'$ be its log canonical model, guaranteed by the log Minimal Model Program. The divisor $K_{X'} + B' + X'_0$ is f'-ample. Inversion of Adjunction – applied in the opposite direction

now – guarantees that the central fiber (X'_0, B'_0) has semi log canonical singularities. Finally, since (X_U, B_U) has log canonical singularities, outside the central fiber the log canonical model of $(\widetilde{X}'_U, \widetilde{B}'_U + \sum \widetilde{E}'_{i,U})$ coincides with $(X_U, B_U) \times_U U'$. So we obtained the desired extension. □

3. Surfaces

The situation with the moduli spaces of surfaces is as follows. The broad outline has been understood for a long time, see [39], [34], [4], but answers to several thorny technical questions have been published only recently. With these technical questions resolved, for any fixed projective scheme Y one can construct the moduli space of stable maps $f\colon (X, B) \to Y$ with B empty or reduced (i.e. with all $b_j = 1$), as a projective scheme. For the arbitrary coefficients b_j, one faces the difficulties with subschemes B_j acquiring embedded components (an example due to Hassett shows that this really happens), and the technical details of the solution are yet to be published. We now give a brief overview.

Definition of the moduli functor. We choose a triple of positive rational numbers $C = (C_1, C_2, C_3)$ and a positive integer N. We also fix a very ample sheaf $\mathcal{O}_Y(1)$ on Y. Then the basic moduli functor $M_{C,N}$ associates to every Noetherian scheme S over a base scheme the set $M_{C,N}(S)$ of maps $f\colon (X, B) \to Y \times S$ with the following properties:

1. X and B_j are flat schemes over S.
2. The double dual $\mathcal{L}_N(X/S) = \left(\omega_{X/S}^{\otimes N} \otimes \mathcal{O}_X(NB)\right)^{**}$ is an invertible sheaf on X, relatively ample over $Y \times S$.
3. For every geometric fiber, $(K_{X_s} + B_s)^2 = C_1$, $(K_{X_s} + B_s)H_s = C_2$, and $H_s^2 = C_3$, where $\mathcal{O}_X(H) = f^*\mathcal{O}_Y(1)$.

Kollár suggested a different moduli functor, of families for which the formation of the sheaves $\mathcal{L}_N(X/S)$ commutes with arbitrary base changes $S' \to S$, i.e.

$$\mathcal{L}_N(X \times_S S'/S') = \phi^*\mathcal{L}_N(X/S)$$

for *all* sheaves $\mathcal{L}_N$ for which NB is integral (e.g., all $N \in \mathbb{Z}$ if B is reduced).

Boundedness. Boundedness means that for any stable map over an algebraically closed field, with fixed invariants C_1, C_2, C_3, there exists N such that the sheaf $L_N = \mathcal{O}_X(N(K_X + B))$ is invertible. Then it is easy to prove that for a fixed multiple M of N the sheaf L_M is very ample with trivial higher cohomologies.

For surface pairs with fixed b_j boundedness was proved in [2], see also [14] for a somewhat simpler, and effective version. For the stable maps it was proved in [4].

(We also note that Karu [30] proved boundedness for *smoothable* stable varieties of dimension d assuming Minimal Model program in dimension $d+1$.)

Local closedness. This means that for every family of pairs $f\colon (X, B) \to Y \times S$, with fibers not assumed to be stable pairs, there exists a locally closed subscheme $U \to S$ with the following universal property: For every $S' \to S$, the pullback family represents an element of $M_N(S')$ if and only if $S' \to S$ factors though U. An important case of this statement from which the general case follows, was established in [28].

Construction of the moduli space. Let $f\colon (X, B) \to Y \times S$ be a family of stable maps over S. By boundedness, for some fixed multiple M of N, the sheaf $\pi_*\mathcal{L}_M$ is locally free, so it can be trivialized on an open cover $S = \cup S_i$. With such trivializations chosen, the graphs of the maps $f_i = f|_{S_i}$ are closed subschemes of $\mathbb{P}^n \times Y$, and so represent a collection of S_i-points of the Hilbert scheme $\mathrm{Hilb}_{\mathbb{P}^n \times Y, p}$, for an easily computable Hilbert polynomial p. For a different choice of trivializations, the points differ by the action of $\mathrm{PGL}_{n+1}(S_i)$.

By local closedness, there exist a locally closed subscheme $U \to \mathrm{Hilb}_{\mathbb{P}^n \times Y, p(x)}$ such that the above S_i-points of the Hilbert scheme are S_i-points of U. Vice versa, every morphism $S \to U$ gives a family of stable maps over S.

It follows that the moduli functor is the quotient functor U/PGL_{n+1}. The separatedness of the moduli functor implies that the PGL_{n+1}-action is proper. Then the quotient exists as an algebraic space by applying either [36] or [31]. It is a proper algebraic space because the moduli functor is proper.

Projectivity of the moduli space. Kollár [34] provided a general method for proving projectivity of complete moduli spaces. It applies in this situation with minor modifications. In particular, the moduli space is a scheme.

We note that the quasiprojectivity of the open part corresponding to arbitrary-dimensional polarized varieties with canonical singularities was proved by Viehweg [48] by using methods of Geometric Invariant Theory.

A floating coefficient version. Hacking [22] constructed a moduli space for the stable pairs $(\mathbb{P}^2, (3/d+\varepsilon)B)$, where B is a plane curve of degree d, and their degenerations.

Other special surfaces. Other papers treating special cases include [1], [25], [47].

4. Toric and spherical varieties

In terms of Definition 1.1, this case corresponds to the pairs $(X, \Delta + \varepsilon B)$, with a floating coefficient. The following very easy statement is the main bridge connecting the log Minimal Model Program and stable pairs with the combinatorics of toric varieties.

We fix a multiplicative torus $T = \mathbb{G}_m^r$. Below, *toric variety* means a normal variety with T-action and an open T-orbit; no special point is chosen (as opposed to a torus embedding).

Lemma 4.1 ([3]). *Let X be a projective toric variety, Δ be the complement of the main T-orbit, and B be an effective $\mathbb{Q}$-Cartier divisor. Then the pair (X, Δ) has log canonical singularities, and the pair $(X, \Delta + \varepsilon B)$ with effective divisor B and $0 < \varepsilon \ll 1$ is stable iff B is an ample Cartier divisor which does not contain any T-orbits.*

Proof. For any toric variety one has $K_X + \Delta = 0$. In particular, we can apply this to X and to a toric resolution of singularities $\pi : \widetilde{X} \to X$. The divisor $\widetilde{\Delta}$, the union of $\pi_*^{-1}\Delta$ and the exceptional divisors E_i, is a normal crossing divisor. But then the formula $K_{\widetilde{X}} + \widetilde{\Delta} = 0 = \pi^*(K_X + \Delta)$ says that the discrepancies a_i in the formula $K_{\widetilde{X}} + \pi_*^{-1}\Delta = \pi^*(K_X + \Delta) + \sum a_i E_i$ all equal -1.

For the pair $(X, \Delta + \varepsilon B)$ to be stable, B must be $\mathbb{Q}$-Cartier and ample, since $K_X + \Delta + \varepsilon B = \varepsilon B$. By continuity of discrepancies of $(X, \Delta + \varepsilon B)$ in ε, we see that the latter pair is log canonical iff $\pi^* B$ does not contain any irreducible components of $\widetilde{\Delta}$, i.e. the closures of proper T-orbits on $\widetilde{X}$. Equivalently, B should not contain any T-orbits. Finally, any effective Weil divisor not containing a T-orbit is Cartier. □

Definition 4.2. Let X be a variety with T-action, and $B \subset X$ be an effective Cartier divisor. The variety X, resp. the pair (X, B) is called a *stable toric variety* (resp. *stable toric pair*) if the following three conditions are satisfied:

1. *on singularities:* X is seminormal (resp. and B does not contain any T-orbits),
2. *on group action:* isotropy groups T_x are subtori (so connected and reduced), and there are only finitely many orbits,
3. *numerical:* (resp. the divisor B is ample).

A *family of stable toric pairs* is a proper flat morphism $f : (X, B) \to S$, where X is a scheme endowed with an action of $T_S := T \times S$, with a relative Cartier divisor B, so that every geometric fiber is a stable toric pair. We will denote the invertible sheaf $\mathcal{O}_X(B)$ by L.

A polarized T-variety (X, L) is *linearized* if X is projective, and the sheaf L is provided with a T-linearization.

We see that a pair (X, B) with a toric variety X is a stable toric pair iff the pair $(X, \Delta + \varepsilon B)$ is a stable pair in the sense of Definition 1.1. We also note that the boundary Δ is determined by the group action, and so can be omitted.

One proves rather easily that a linearized stable toric variety is a union of (normal) polarized toric varieties (X_i, L_i) which, as it is well known [44], correspond to lattice polytopes Q_i. In this way, one obtains a *complex of polytopes* $\mathcal{Q} = (Q_i)$, and X is

glued from the varieties X_i combinatorially in the same way as the topological space $|\mathcal{Q}|$ is glued from Q_i. The complex $\mathcal{Q}$ comes with a *reference map* $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}$, where M is the character group of T, identifying each cell Q_i with a lattice polytope. The pair $(|\mathcal{Q}|, \rho)$ is called the *type* of a stable toric variety.

A section $s \in H^0(X, L)$ with $B = (s)$ gives a collection of sections

$$s_i = \sum s_{i,m} e^m \in H^0(X_i, L_i) = \bigoplus\nolimits_{m \in Q \cap M} k e^m.$$

For each polytope Q_i this gives a subset $C_i = \{m \mid s_{i,m} \neq 0\}$ and, since B does not contain any T-orbits, one must have $\operatorname{Conv} C_i = Q_i$. This defines a *complex of marked polytopes* $(\mathcal{Q}, \mathcal{C})$.

4.3. All stable toric varieties X (resp. pairs (X, B)), are classified, up to an isomorphism, by the following data:

1. A complex of polytopes $\mathcal{Q}$ with a reference map $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}$ (resp. a complex of marked polytopes $(\mathcal{Q}, \mathcal{C})$ with a reference map).

2. An element of a certain cohomology group which we briefly describe.

For each polytope $Q_i \in \mathcal{Q}$, let $\widetilde{M}_i \subset \widetilde{M} = \mathbb{Z} \times M$ be the saturated sublattice of $\widetilde{M}$ generated by $(1, Q_i)$, and let $\widetilde{T}_i = \operatorname{Hom}(\widetilde{M}_i, \mathbb{G}_m)$ be the corresponding torus. The collection of stalks $\{\widetilde{T}_i\}$ defines the sheaf $\underline{\widetilde{T}}$ on the complex $\mathcal{Q}$. Then the set of isomorphism classes of polarized stable toric varieties is simply $H^1(\mathcal{Q}, \underline{\widetilde{T}})$, and each of them has automorphism group $H^0(\mathcal{Q}, \underline{\widetilde{T}})$.

Similarly, one defines the sheaf $\underline{\widehat{C}} = \operatorname{Hom}(\mathcal{C}, \mathbb{G}_m)$ with the stalks $\operatorname{Hom}(C_i, \mathbb{G}_m)$ in which the sections s_i live. The natural sheaf homomorphism $\underline{\widetilde{T}} \to \underline{\widehat{C}}$ gives a homomorphism of cochain complexes $\phi\colon C^*(\underline{\widetilde{T}}) \to C^*(\underline{\widehat{C}})$. Then the first cohomology of the cone complex $\operatorname{Cone}(\phi)$ is the set of isomorphism classes of stable toric pairs of type $(\mathcal{Q}, \mathcal{C})$, and the zero cohomology gives the automorphism groups of the pairs. These automorphism groups are finite.

The following lemma also goes back to [3].

Lemma 4.4. *Suppose that the topological space $|\mathcal{Q}|$ is homeomorphic to a manifold with boundary. Let (X, B) be a stable toric pair in the sense of Definition* 4.2. *Let Δ be the reduced divisor corresponding to the boundary of $|\mathcal{Q}|$. Then the pair $(X, \Delta + \varepsilon B)$ is stable in the sense of Definition* 1.1.

Proof. One proves that with the above assumption on $|\mathcal{Q}|$ the variety X is Cohen–Macaulay, a fortiori, satisfies S2, and has only simple crossings in codimension 1 (each component of the double locus corresponds to a codimension-1 polytope in $\mathcal{Q}$ which is a face of two maximal-dimensional polytopes). The normalization of $(X, \Delta + \varepsilon B)$ with the double locus added is the disjoint union of toric pairs $(X_i, \Delta_i + \varepsilon B_i)$, which are log canonical by Lemma 4.1. Hence, $(X, \Delta + \varepsilon B)$ is semi log canonical. Moreover,

$$\nu^*(K_X + \Delta + \varepsilon B)\,|_{X_i} = K_{X_i} + \Delta_i + \varepsilon B_i,$$

so the divisor $K_X + \Delta + \varepsilon B$ is ample. □

We would like to mention the following essential facts: Higher cohomology groups of positive powers L^d vanish. The moduli functor of stable toric pairs is proper, i.e. every one-parameter family has at most one limit, and the limit always exists after a finite base change $(S', 0) \to (S, 0)$. The limit of a family of pairs of type $\mathcal{Q}$ corresponds to a complex $\mathcal{Q}'$ such that $|\mathcal{Q}'| = |\mathcal{Q}|$, and $\mathcal{Q}'$ is obtained from $\mathcal{Q}$ by a *convex* subdivision.

Recall that a subdivision of a single polytope Q is *convex* if it is the projection of the lower envelope of several points $\big(m, h(m)\big)$ where m are some points with $\mathrm{Conv}(m) = Q$, and $h\colon \{m\} \to \mathbb{R}$ is an arbitrary function, called *height function*. This was generalized in [6] to convex subdivisions of a polytopal complex $\mathcal{Q}$ by requiring that the height functions of two polytopes Q_1, Q_2 differ by a linear function on $Q_1 \cap Q_2$.

The stable toric variety X is *multiplicity-free* if the reference map $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}$ is injective. We will restrict ourselves to this case for the rest of this section.

Theorem 4.5 ([6]). *The functor of stable toric pairs has a coarse moduli space M over $\mathbb{Z}$. It is a disjoint union of subschemes $M_{|\mathcal{Q}|}$, each of them projective. Each moduli space $M_{|\mathcal{Q}|}$ has a natural stratification with strata corresponding to subdivisions of $|\mathcal{Q}|$ into lattice polytopes.*

When $|\mathcal{Q}| = Q$ is a polytope, the moduli space M_Q contains an open subset U_Q which is the moduli space of pairs (X, B) with a toric variety X. The closure of U_Q is an irreducible component of M_Q. The strata in this closure correspond to convex subdivisions of Q, and the normalization of $\overline{U}_Q$ coincides with the toric variety for the secondary polytope of $(Q, Q \cap M)$.

Rather than relying on the methods of the Minimal Model Program, the proof is rather direct. To each family $(T_S \curvearrowright X, B) \to S$ we can associate the graded algebra $R(X/S, L) = \bigoplus_{d \geq 0} \pi_* L^d$, multigraded by M due to the T_S-action, and multiplicity-free by the assumptions on the fibers, with a section s, an equation of B. Then the moduli of stable toric pairs is equivalent to the moduli of algebras (R, s) with a section, and the latter is rather straightforward.

We note that the faces of the secondary polytope of $(Q, Q \cap M)$ (see [20] for the definition) are in bijection with the convex subdivisions of Q.

For some polytopes Q the moduli space M_Q does indeed have several irreducible components. The extra components always appear when there exists a non-convex subdivision of Q into lattice polytopes.

Another situation where the extra components are guaranteed is when the stratum for a particular convex subdivision has higher dimension in M_Q than it does in the toric variety for the secondary polytope. Both can be computed effectively: the latter is the codimension of the corresponding cone in the normal fan of the secondary polytope, and for M_Q it is the dimension of the cohomology group describing the gluing, as in 4.3.

On the other hand, the following dual point of view turns out to be very important.

Definition 4.6. Let Y be a projective scheme. A *variety over* $\boldsymbol{Y}$ is a reduced, but possibly reducible, projective variety X together with a finite morphism $f\colon X \to Y$. A *family of varieties over* $\boldsymbol{Y}$ is a finite morphism $f\colon X \to Y \times S$ such that $X \to S$ is flat, and every geometric fiber is a variety over Y, as above.

If G is an algebraic group acting linearly on $Y \subset \mathbb{P}^n$, then a G-variety (resp. family) over Y is a morphism $f\colon X \to Y$ as before which, in addition, is G-equivariant.

Lemma 4.7. *Families of stable toric pairs of type $|\mathcal{Q}|$ are in a natural bijective correspondence with families of stable toric varieties over $\mathbb{P}^n$ with the homogeneous coordinates x^m, $m \in M \cap |\mathcal{Q}|$, on which T acts with the characters m.*

Proof. Indeed, the data for both the morphism to $\mathbb{P}^n$ and the divisor B not containing any T-orbits is the same: working locally over $S = \operatorname{Spec} A$, it is a collection $(c_m \in A)$ such that $c_m(s) \neq 0$ for every *vertex* m of a polytope Q_i corresponding to the fiber X_s. □

Remark 4.8. We note that there exists another moduli space closely related to our moduli space M of stable toric varieties: it is the toric Hilbert scheme $\operatorname{Hilb}^T_{\mathbb{P}^n}$ [46], [24] parameterizing *subschemes* $X \subset \mathbb{P}^n$ corresponding to the multiplicity-free multigraded algebras. So, geometrically what we have done is the following: we replaced *closed subschemes* of Y by *reduced* varieties X with a *finite morphism to* Y.

Indeed, this is a general situation, and in [13] such a universal substitute for the Hilbert scheme is constructed in general, without the multiplicity-free assumption (or group action). Reduced varieties with a finite morphism to a scheme Y are called *branchvarieties of* $\boldsymbol{Y}$, to contrast with subvarieties or subschemes of Y.

Over a field of characteristic zero, the above picture can be generalized to stable spherical varieties over Y. Recall that if G is a connected reductive group then a G-variety X is called *spherical* if it is normal and a Borel subgroup of G has an open orbit. One motivation for considering spherical varieties is the following important finiteness property: a G-variety is spherical iff any G-variety birationally isomorphic to it has only finitely many G-orbits.

A polarized G-linearized variety (X, L) is spherical iff it is normal and the algebra $R(X, L) = \bigoplus_{d\geq 0} H^0(X, L^d)$ is multiplicity free, i.e. when it is written as a direct sum over the irreducible representations $V_{d,\lambda}$ of the group $\widetilde{G} = \mathbb{G}_m \times G$, each multiplicity is 1 or 0.

One important difference between the toric and spherical cases is that the spherical varieties are *not* completely classified. For any homogeneous spherical variety G/H, its normal G-embeddings correspond to colored fans. However, the homogeneous spherical varieties G/H are currently only classified in types A and D, [41], [17].

For a polarized spherical variety (X, L) one can define its moment polytope [18] which, when working over $\mathbb{C}$, coincides with the moment polytope of X as a Hamiltonian variety. Which polytopes may appear as moment polytopes is not known.

However, it is known that the set of moment polytopes contained in any bounded set is finite [11], [12]. Together with the boundedness results of [13], this implies that the set of moment polytopes of polarized stable spherical varieties (X, L) with a fixed Hilbert polynomial is finite.

Definition 4.9. A *stable spherical variety over* a G-variety $\boldsymbol{Y} \subset \mathbb{P}^n$ is a G-variety over Y such that the G-module $R(X, L)$, $L = f^*\mathcal{O}_Y(1)$ is multiplicity-free,

Similarly, a *family of stable spherical varieties over* $\boldsymbol{Y}$ is a proper family of G-varieties $f\colon G_S \curvearrowright X \to Y \times S$ over Y such that, denoting $L = f^*\mathcal{O}_{Y\times S}(1)$, one has

$$R(X/S, L) = \bigoplus_{d\geq 0} \pi_*(X, L^d) = \bigoplus_\lambda V_{d,\lambda} \otimes F_{d,\lambda},$$

where $V_{d,\lambda}$ are the irreducible $(\mathbb{G}_m \times G)$-representations and $F_{d,\lambda}$ are locally free sheaves of rank 1 or 0.

There is an equivalent more geometric definition in terms of gluing from spherical varieties. However, as was noted above, the structure of the "building blocks", i.e. ordinary spherical varieties, is a little mysterious.

Theorem 4.10 ([12]). *The functor of stable spherical pairs over Y has a coarse moduli space M_Y over $\mathbb{Q}$. It is a disjoint union of projective schemes.*

As in the toric case, one can define a *stable spherical pair* (X, B). However, when G is not a torus, this turns out to be a very special case of stable maps. The analogue of Lemma 4.7 in the spherical case is the following:

Lemma 4.11 ([12], Prop.3.3.2). *Families of stable spherical pairs of type $|\mathcal{Q}|$ are in a natural bijective correspondence with families of stable spherical varieties over $\mathbb{P}\big(\oplus \operatorname{End}(V_m)\big)$, where m go over the weights in $|\mathcal{Q}|$.*

One important case where all stable spherical varieties are completely classified is the case of *stable reductive varieties* [9], [10]. Each polarized stable reductive variety corresponds to a complex $\mathcal{Q} = (Q_i)$ of lattice polytopes in $\Lambda_{\mathbb{R}}$, where Λ is the weight lattice of G, and the complex $\mathcal{Q}$ is required to be invariant under the action of Weyl group. Limits of one-parameter degenerations again correspond to convex subdivisions.

5. Abelian varieties

In terms of Definition 1.1, this case corresponds to the pairs $(X, \varepsilon B)$, with a floating coefficient. Here, X is an abelian variety, or more accurately an abelian torsor (i.e. no origin is fixed) or a similar "stable" variety, and B is a theta divisor. But again, Minimal Model Program is not used, and the constructions and proofs of [6] are more direct, using the symmetries of the situation.

A fundamental insight from the Minimal Model Program is that we should be working with abelian torsors with divisors ($B \subset X$) instead of abelian varieties ($0 \in X$), because the former fit into the general settings of Definition 1.1 and the basic construction of Section 2, and the latter do not. The bridge is the following

Lemma 5.1 ([6]). *There is a natural bijective correspondence between principally polarized abelian schemes $(A, \lambda\colon A \to A^t) \to S$ and flat families of abelian torsors $(A \curvearrowright X, B) \to S$ such that B is an effective relative Cartier divisor defining a principal polarization on each geometric fiber.*

For example, if $C \to S$ is a smooth family of curves then $(\mathrm{Pic}^0_{C/S}, \lambda)$ is the principally polarized abelian scheme, and $(\mathrm{Pic}^0_{C/S} \curvearrowright \mathrm{Pic}^{g-1}_{C/S} \supset \Theta_{g-1})$ is the family of abelian torsors. The two families are usually not isomorphic, unless $C \to S$ has a section.

Recall that a *semiabelian variety* is a group variety G which is an extension

$$1 \to T \to G \to A \to 0$$

of an abelian variety by a multiplicative torus. Let $g = \dim G = r + a = \dim T + \dim A$. We will denote the lattice of characters of T by $M_0 \simeq \mathbb{Z}^r$ and reserve $M \simeq \mathbb{Z}^g$ for a certain lattice of which M_0 will be a quotient.

Generalizing directly Definition 4.2, we give the following:

Definition 5.2. Let X be a variety with an action of a semiabelian variety G, and $B \subset X$ be an effective Cartier divisor. The variety X, resp. the pair (X, B) is called a *stable quasiabelian variety* (resp. *stable quasiabelian pair*) if the following three conditions are satisfied:

1. *on singularities:* X is seminormal (resp. and B does not contain any G-orbits),
2. *on group action:* isotropy groups G_x are subtori, and
3. *numerical:* (resp. the divisor B is ample).

A proper flat morphism $f : (G \curvearrowright X, B) \to S$ is called a *family of stable quasiabelian pairs*; here G is a semiabelian group scheme over S, X is a scheme endowed with an action $G \times_S X \to X$, with a relative Cartier divisor B, so that every geometric fiber is a stable quasiabelian pair.

The essential difference with the case of stable toric varieties is that the group variety G may vary, and in particular the torus part T may change its rank.

Intuitively (and this actually works when working over $\mathbb{C}$) a polarized abelian variety should be thought of as a stable toric variety for a *constant* torus that, in terms of the data 4.3, corresponds to a topological space $|\mathcal{Q}| = \mathbb{R}^g/\mathbb{Z}^g$ and an element of a cohomology group describing the gluing, as in 4.3.2; however, the Čech cohomology should be replaced by the group cohomology. A general polarized stable quasiabelian variety should be thought of as a similar quotient of a bigger stable toric variety for a constant torus by a lattice.

Remark 5.3. Namikawa [43] defined SQAVs, or "stable quasiabelian varieties" not intrinsically but as certain limits of abelian varieties. They are different from our varieties in several respects. In particular, some of Namikawa's varieties are not reduced, and it is not clear if they vary in flat families. The above definition also includes varieties which are not limits of abelian varieties.

The varieties of Definition 5.2 were called stable semiabelic varieties in [6]. Since this was a somewhat awkward name, I am reverting to the old name.

A stable quasiabelian pair is *linearized* if the sheaf $L = \mathcal{O}_X(B)$ is provided with a T-linearization (*not* G-linearization!) The first step is to classify the linearized varieties, which is quite easy:

Theorem 5.4. *A linearized stable quasiabelian variety* $(G \curvearrowright X, L)$ *is equivalent to the following data:*

1. (*toric part*) *a linearized stable toric variety* (X_0, L_0) *for* T*, and*
2. (*abelian part*) *a polarized abelian torsor* (X_1, L_1) *for* A.

The variety X *is isomorphic to the twisted product*

$$X = X_0 \times^T G = (X_0 \times G)/T, \quad T \text{ acting by } (t, t^{-1})$$

and there is a locally trivial fibration $X \to X_1$ *with fibers isomorphic to* X_0*. Each closed orbit of* $G \curvearrowright X$ *with the restriction of* L *is isomorphic to the pair* (X_1, L_1).

5.5. Hence, the linearized stable quasiabelian varieties (X, L) (resp. pairs (X, B)) over a field $k = \bar{k}$ are easy to classify, and are described by the following data:

1. A complex of polytopes $\mathcal{Q}_0$ with a reference map $\tilde{\rho}_0 \colon |\mathcal{Q}_0| \to M_{0,\mathbb{R}}$.
2. A polarized abelian torsor (X_1, L_1), which is equivalent to a polarized abelian variety $(A, \lambda \colon A \to A^t)$.
3. A semiabelian variety G, which is equivalent to a homomorphism $c \colon M_0 \to A^t$.
4. A certain cohomology group, very similar to the one in 4.3, describing the gluing. The only difference is that in this case the sheaves, instead of tori, have coefficients in certain $\mathbb{G}_m$-torsors.

The topological space $|\mathcal{Q}_0|$ has dimension $\dim X_0 \leq \dim T = r$, the toric rank of G. The analogy with stable toric varieties becomes even stronger when we associate to (X, L) a *cell* complex $\mathcal{Q}$ of dimension $\dim X$. This is done as follows:

The kernel of the polarization map $\lambda \colon A \to A^t$ has order d^2, where d is the degree of polarization, and comes with a skew-symmetric bilinear form. If $\operatorname{char} k \nmid d$, it can be written as $\ker \lambda = H \times \operatorname{Hom}(H, \mathbb{G}_m)$ for a unique finite abelian group H of rank $\leq \dim A$. Write H as a quotient M_1/Γ_1, where $M_1 = \mathbb{Z}^a$ and Γ_1 is a subgroup of finite index.

The cell complex $\mathcal{Q}_1$ we associate to the abelian torsor (X_1, L_1) consists of one cell $M_{1,\mathbb{R}}/\Gamma_1 \simeq \mathbb{R}^a/\mathbb{Z}^a$, a real torus of dimension a. Let $M = M_0 \times M_1$ and $\Gamma = \Gamma_1$. Then the cell complex associated to the linearized pair (X, L) is $\mathcal{Q} = \mathcal{Q}_0 \times \mathcal{Q}_1$. It comes with a reference map $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}/\Gamma_1 \simeq \mathbb{R}^g/\mathbb{Z}^a$.

It turns out, however, that if $(T \curvearrowright X, B)$ is a degeneration of abelian varieties with $T \neq 1$, the invertible sheaf L is *never* linearized. On the other hand, there exists an infinite *algebraic* cover $f\colon \widetilde{X} \to X$ such that the pullback $\widetilde{L} = f^*L$ is linearized in a canonical way:

Theorem 5.6 ([6]). *Let G be a semiabelian group scheme over a connected scheme S, and assume that G is a global extension $1 \to T \to G \to A \to 0$ of an abelian scheme A by a split torus T. Then a family of stable quasiabelian pairs $(G \curvearrowright X, B) \to S$ is equivalent to a family of* linearized *stable quasiabelian pairs $(G \curvearrowright \widetilde{X}, \widetilde{B}) \to S$ whose fibers are only* locally of finite type*, with a compatible* free in Zariski topology *action of $M_0 = \mathbb{Z}^r$ so that $(X, B) = (\widetilde{X}, \widetilde{B})/M_0$.*

Moreover, there exists a subgroup $\Gamma_0 \simeq \mathbb{Z}^{r'}$, $r' \leq r$, of M_0 such that $\widetilde{X}$ is a disjoint union of $[M_0 : \Gamma_0]$ copies of a connected scheme $\widetilde{X}'$, and $(X, B) = (\widetilde{X}', \widetilde{B}')/\Gamma_0$.

A polarized toric variety (X, L) provides a trivial case of this theorem: in this case $\Gamma_0 = 0$, and X is the disjoint union of $M_0 \simeq \mathbb{Z}^r$ copies of X, one for each possible T-linearization of the sheaf L.

A less trivial, but equally familiar example is the rational nodal curve, which is a quotient $\widetilde{X}/\mathbb{Z}$ of an infinite chain of $\mathbb{P}^1$s by $M_0 \simeq \mathbb{Z}$. Mumford [42] constructed a number of degenerations of abelian varieties which are such infinite quotients. So the above statement may be considered to be a precise inverse of Mumford's construction.

The scheme $\widetilde{X}'$ can be written in the form $\widetilde{X}_0 \times^T G$, where $\widetilde{X}_0$ is a linearized scheme locally of finite type. Then locally $\widetilde{X}_0$ is isomorphic to a linearized stable toric variety. This defines a locally finite complex of polytopes $\widetilde{\mathcal{Q}}_0$ with a reference map $\tilde{\rho}_0\colon |\widetilde{\mathcal{Q}}_0| \to M_{0,\mathbb{R}}$. Moreover, $\widetilde{\mathcal{Q}}_0$ has a Γ_0-action with only finitely many orbits, and $\tilde{\rho}_0$ is Γ_0-invariant. This can be summed up by saying that each stable quasiabelian pair defines a finite complex of polytopes $\rho_0\colon \mathcal{Q}_0 = \widetilde{\mathcal{Q}}_0/\Gamma_0 \to M_{0,\mathbb{R}}/\Gamma_0 \simeq \mathbb{R}^r/\mathbb{Z}^{r'}$.

Again, we can add to this the abelian part, and obtain a complex of dimension $\dim X$. As above, the abelian part gives a one-cell complex $\mathcal{Q}_1 = M_{1,\mathbb{R}}/\Gamma_1 \simeq \mathbb{R}^a/\mathbb{Z}^a$. We set $\mathcal{Q} = \mathcal{Q}_0 \times \mathcal{Q}_1$, $M = M_0 \times M_1$, and $\Gamma = \Gamma_0 \times \Gamma_1$. Then $\mathcal{Q}$ is a finite cell complex, and it comes with a reference map $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}/\Gamma \simeq \mathbb{R}^g/\mathbb{Z}^{a+r'}$.

The topological space $\rho\colon |\mathcal{Q}| \to M_{\mathbb{R}}/\Gamma$ together with the reference map is the *type of a stable quasiabelian variety* (X, L), resp. of a pair (X, B). We say that the type is *injective* if the reference map ρ is injective. In this case it can be shown that the type is constant in connected families, and so we can talk about moduli spaces $M_{|\mathcal{Q}|}$.

5.7. The classification of isomorphism classes in the stable toric case 4.3 and linearized stable semiabelian case 5.5 can be translated to this most general case almost verbatim.

The most important of the moduli spaces $M_{|\mathcal{Q}|}$ are the ones containing abelian torsors of degree d, defining a polarization $\lambda\colon A \to A^t$ with $\ker\lambda \simeq H \times \mathrm{Hom}(H, \mathbb{G}_m)$. In this case, the type is a real torus $M_{\mathbb{R}}/\Gamma \simeq \mathbb{R}^g/\mathbb{Z}^g$, where $\Gamma \subset M$ is a sublattice with $M/\Gamma \simeq H$. One observes that this *real torus with a lattice structure* is an analogue of a lattice polytope Q in the stable toric case.

A cell subdivision of $|\mathcal{Q}|$ in this case is the same as a Γ-periodic subdivision of $M_{\mathbb{R}}$ which is a pullback of a Γ_0-periodic subdivision of $M_{0,\mathbb{R}}$ into lattice polytopes with vertices in M_0.

One proves that the moduli functor of stable quasiabelian pairs of injective type is proper, and that one-parameter degenerations correspond to suitably understood convex subdivisions of $|\mathcal{Q}|$.

A Γ-periodic subdivision of $M_{\mathbb{R}}$ is *convex* if it is the projection of the lower envelope of the points $(m, h(m))$, where m goes over $M \simeq \mathbb{Z}^g$, and $h\colon M \to \mathbb{R}$ is a function of the form

$$h(m) = (\text{positive semidefinite quadratic form}) + r(m \bmod \Gamma),$$

where $r\colon M/\Gamma \to \mathbb{R}$ is a function defined on the finite set of residues.

In particular, the principally polarized case corresponds to $\Gamma = M = \mathbb{Z}^g$. The convex subdivisions of $\mathbb{R}^g/\mathbb{Z}^g$ in this case are the classical *Delaunay decompositions*, that appeared in [50]. A detailed combinatorial description of one-parameter degenerations of principally polarized abelian varieties from the present point of view, in which Delaunay decompositions naturally appear, was given in [15].

By analogy, when $\Gamma \subset M$ is a sublattice of finite index, we call the convex subdivisions of $M_{\mathbb{R}}/\Gamma$ *semi-Delaunay decompositions*.

We will denote the moduli spaces appearing in this case by $\overline{\mathrm{AP}}_{g,H}$, resp. $\overline{\mathrm{AP}}_g$ in the principally polarized case; AP stands for *abelian pairs*.

Theorem 5.8. *For each of the types $|\mathcal{Q}| = M_{\mathbb{R}}/\Gamma \simeq \mathbb{R}^g/\mathbb{Z}^g$, $|M/\Gamma| = d$, the functor of stable quasiabelian pairs of type $|\mathcal{Q}|$ over $\mathbb{Z}[1/d]$ has a coarse moduli space $\overline{\mathrm{AP}}_{g,H}$, a proper algebraic space over $\mathbb{Z}[1/d]$. The moduli space $\overline{\mathrm{AP}}_{g,H}$ has a natural stratification with strata corresponding to subdivisions of $|\mathcal{Q}|$ modulo symmetries of (M, Γ).*

The moduli space $\overline{\mathrm{AP}}_{g,H}$ contains an open subset $U_{|\mathcal{Q}|}$ of dimension $\frac{g(g+1)}{2} + d - 1$ which is the moduli space of abelian torsors (X, B) defining a polarization of degree d. The closure of $U_{|\mathcal{Q}|}$ is an irreducible component of $\overline{\mathrm{AP}}_{g,H}$. The strata in this closure correspond to semi-Delaunay subdivisions.

In particular, one has the following:

Theorem 5.9. *For $|\mathcal{Q}| = M_{\mathbb{R}}/M \simeq \mathbb{R}^g/\mathbb{Z}^g$, the functor of stable quasiabelian pairs of type $|\mathcal{Q}|$ has a coarse moduli space $\overline{\mathrm{AP}}_g$, a proper algebraic space over $\mathbb{Z}$. The moduli space $\overline{\mathrm{AP}}_g$ has a natural stratification with strata corresponding to $\mathbb{Z}^g$-periodic subdivisions of $\mathbb{R}^g$, pullbacks of $\mathbb{Z}^r$-periodic subdivisions of $\mathbb{R}^r$ into lattice polytopes with the set of vertices equal to the lattice $\mathbb{Z}^r$ of periods, modulo* $\mathrm{GL}(g, \mathbb{Z})$.

The moduli space $\overline{\mathrm{AP}}_g$ contains an open subset which is the moduli space A_g of principally polarized abelian varieties. The closure of A_g is an irreducible component of $\overline{\mathrm{AP}}_g$. The strata in this closure correspond to Delaunay subdivisions, and the normalization of $\bar{A}_g$ coincides with the toroidal compactification $\bar{A}_g^{\mathrm{vor}}$ for the second Voronoi fan. This toroidal compactification is projective.

The proof of the theorems exploits the connection with the toric case in the following way. Although the toric rank in a family of stable quasiabelian varieties may change, in an infinitesimal family it does not. Hence, Theorem 5.6 together with the toric methods give the deformation and obstruction theory for the moduli functor. The moduli spaces then can be constructed by using Artin's method [16].

The (locally closed) cones of the second Voronoi fan consist of positive semidefinite quadratic forms that define the same Delaunay decomposition. Thus, we see that this fan, introduced by Voronoi in [50], is the precise infinite periodic analogue of (the normal fan of) the secondary polytope, and predates it by about 80 years.

Starting with $g = 4$, the moduli spaces $\overline{\mathrm{AP}}_g$ do indeed have several irreducible components (as do the moduli spaces of stable toric varieties). The extra components always appear when there exists a non-Delaunay $\mathbb{Z}^r$-periodic subdivision of $\mathbb{R}^r$ into lattice polytopes with vertices in the same $\mathbb{Z}^r$, with $r \leq g$.

Another situation where the extra components are guaranteed is when the stratum for a particular Delaunay decomposition has higher dimension in $\overline{\mathrm{AP}}_g$ than it does in $\bar{A}_g^{\mathrm{vor}}$. Both can be computed effectively: for the Voronoi compactification it is the codimension of a cone in the 2nd Voronoi fan, and for $\overline{\mathrm{AP}}_g$ it is the dimension of the cohomology group describing the gluing, as in 4.3, 5.5, 5.7. See [5] for more on this.

One important construction involving moduli spaces of abelian varieties is the Torelli map $M_g \to A_g$ which associates to a smooth curve C of genus g its Jacobian, a principally polarized abelian variety $(A, \lambda\colon A \to A^t)$ of dimension g. Combinatorially, it was understood by Mumford (see [43]) that the Torelli map can be extended to a morphism from the Deligne–Mumford compactification $\bar{M}_g$ to $\bar{A}_g^{\mathrm{vor}}$, and this was the original motivation for considering the second Voronoi fan in [43].

The moduli interpretation of the extended morphism $\bar{M}_g \to \overline{\mathrm{AP}}_g$ was given in [7]. To a nonsingular curve C, it associates the pair $(\mathrm{Pic}^0 C \curvearrowright \mathrm{Pic}^{g-1} C, \Theta)$, and to a stable curve, the stable quasiabelian pair $(\mathrm{Pic}^0 C \curvearrowright \mathrm{Jac}_{g-1} C, \Theta)$, where $\mathrm{Jac}_{g-1} C$ is the moduli stable of semistable rank 1 sheaves on C of degree $g-1$, and Θ is the divisor corresponding to sheaves with sections.

The $\mathbb{Z}^g$-periodic cell decompositions corresponding to the image of $\bar{M}_g$ have a simple description. First of all, they are not arbitrary but are given by subdividing $\mathbb{R}^r$ (and by pullback, $\mathbb{R}^g$) by systems of parallel hyperplanes $\{l_i(m) = n \in \mathbb{Z}\}$. The condition that the set of vertices of the polytopes in $\mathbb{R}^r$ cut out by the hyperplanes is the same lattice of periods $\mathbb{Z}^r$ is equivalent to the condition that $\{l_i \in M_0^\vee\}$ is a *unimodular system of vectors*, i.e. every $(r \times r)$-minor is 0, 1 or -1. Another name used for unimodular systems of vectors is *regular matroid*, see [45].

In these terms, the answer to "combinatorial Schottky" or "tropical Schottky" problem is the following: the strata in the image of $\overline{M}_g$ correspond to special matroids that are called *cographic*. If $\mathcal{G}$ is a graph then its cographic subdivision is the $H_1(\mathcal{G}, \mathbb{Z})$-periodic subdivision of $H_1(\mathcal{G}, \mathbb{R})$ obtained by intersecting H_1 with $C_1(\mathcal{G}, \mathbb{R})$ divided into standard Euclidean cubes.

This theory was extended to the degenerations of Prym varieties in [8], [49], and many non-cographic matroids appear there. For example, using the above theory Gwena [21] described some of the degenerations of intermediate Jacobians of cubic 3-folds, including a particular one that corresponds to a very symmetric regular matroid R_{10} which is neither cographic, nor graphic.

In conclusion, we would like to mention one other important motivation from the Minimal Model Program for looking at stable abelian pairs: by a theorem of Kollár [37], if A is a principally polarized variety then the pair (A, Θ) has log canonical singularities.

6. Grassmannians

In Section 4 we defined stable toric (and spherical) varieties over a projective G-scheme $Y \subset \mathbb{P}^n$. The corresponding moduli spaces $M_{Y,Q}$ are projective. What does one get by looking at some particular varieties Y? One nice case that has many connections with other fields is the case when $Y \subset \mathbb{P}^n$ is a grassmannian with its Plücker embedding and the group is the multiplicative torus.

Let $E = E_1 \oplus \cdots \oplus E_n$ be a linear space with the coordinate-wise action by the torus $T = \mathbb{G}_m^n$ with character group $M = \mathbb{Z}^n$. (Dimensions of E_i are arbitrary.) Let r be a positive integer, and

$$i : Y = \mathrm{Gr}(r, E) \hookrightarrow \mathbb{P}(\Lambda^r E)$$

be the grassmannian variety of r-dimensional subspaces of E with its Plücker embedding.

For each collection of 2^n nonnegative integers $\underline{d} = (d_I \mid I \subset \{1, \ldots, n\})$, the *thin Schubert cell* is defined to be the locally closed subscheme of $\mathrm{Gr}(r, E)$

$$\mathrm{Gr}_{\underline{d}} = \left\{ V \subset E \mid \mathrm{rank}(V \cap \textstyle\bigoplus_{i \in I} E_i) = d_I \right\}.$$

(Some of these may be empty; one necessary condition for non-emptiness is the inequality $d_{I \cap J} + d_{I \cup J} \geq d_I + d_J$ for all I, J.) There is a subtorus $T_{\underline{d}}$ acting trivially on $\mathrm{Gr}_{\underline{d}}$, and the quotient torus $(T / T_{\underline{d}})$ acts freely.

A *generalized matroid polytope* is a polytope in $\mathbb{R}^n$ defined by the inequalities

$$Q_{\underline{d}} = \left\{ 0 \leq x_i \leq \dim E_i,\ \textstyle\sum_{i=1}^n x_i = r \text{ and } \sum_{i \in I} x_i \geq d_I \text{ for all } I \right\}.$$

The classical *matroid polytopes* are a special case, when all $\dim E_i = 1$.

In [40], Lafforgue constructed certain compactifications of the quotients of thin Schubert cells

$$\mathrm{Gr}_{\underline{d}}/T = \mathrm{Gr}_{\underline{d}}/(T/T_{\underline{d}}).$$

These compactifications have important applications in Langlands program, and they include compactifications of the homogeneous spaces $\mathrm{PGL}_r^{n-1}/\mathrm{PGL}_r$ equivariant with respect to the action by PGL_r (this is the case of all $\dim E_i = r$ and $d_I = 0$ for all $I \neq \{1, \dots, n\}$).

As shown in [12], the main irreducible components of the moduli spaces $M_{\mathrm{Gr}(r,E),Q_{\underline{d}}}$ of stable toric varieties over $\mathrm{Gr}(r, E)$, as in Section 4, coincide with the Lafforgue's compactifications, at least up to normalization.

The connection is as follows. For each point $p \in \mathrm{Gr}_{\underline{d}}$, the normalization of the closure of the orbit $T \cdot p$ defines a T-toric variety $X \to \mathrm{Gr}(r, E)$ for the polytope $Q_{\underline{d}}$. This gives a canonical identification of $\mathrm{Gr}_{\underline{d}}/T$ with an open subset U of the moduli space $M_{\mathrm{Gr}(r,E),Q_{\underline{d}}}$. Since the latter is projective, this gives a moduli compactification of $\mathrm{Gr}_{\underline{d}}/T$.

A case of particular interest is when all $\dim E_i = 1$ and Q is the moment polytope of a generic point $p \in \mathrm{Gr}(r, n)$. In this case $Q_{\underline{d}} = \Delta(r, n)$, a *hypersimplex*. This case was considered by Kapranov in [29] who constructed a compactification he called the Chow quotient, by using the Chow variety.

The moduli space $M = M_{\mathrm{Gr}(r,n),\Delta(2,n)}$ in this case can be interpreted as a compactified moduli space of hyperplane arrangements. This interpretation, with a somewhat folk status, was recorded by Hacking–Keel–Tevelev in [23], along with many new results about this moduli space. We note that the latter paper uses the toric Hilbert scheme, rather than stable toric varieties over Y. But in this case the two points of view coincide, because matroid polytopes, as was observed in [23], are unimodular (by which we mean that the monoid of integral points in the cone over Q is generated by the integral points of Q). As a consequence, stable toric varieties over $\mathrm{Gr}(r, n)$ are actually T-invariant *subschemes* of $\mathrm{Gr}(r, n)$.

We now review this interpretation. Let $p \in \mathrm{Gr}^0(r, n)$ be a generic point, and $X = \overline{T \cdot p}$ be the orbit closure, isomorphic to a (normal) toric variety for the polytope $\Delta(r, n)$. Then $X \to \mathrm{Gr}(r, n)$ can be equivalently interpreted as any of the following:

1. a point in $\mathrm{Gr}^0(r, n)/T$,
2. a point of an open subset $U = U_{\mathrm{Gr}(r,n),\Delta(r,n)}$ of $M_{\mathrm{Gr}(r,n),\Delta(r,n)}$,
3. a point of an open subset U of the toric Hilbert scheme of $\mathrm{Gr}(r, n)$.

Now pick a generic point $e \in \mathbb{A}^n$ and consider the closed subvariety Y_e of X corresponding to the r-dimensional subspaces that contain e. This is Kapranov's *visible contour*. It is easy to see that:

1. For generic e, e' the varieties Y_e and $Y_{e'}$ are isomorphic since they differ by T-action. So we could as well associate one variety Y with each X, up to an isomorphism.

2. As the line $\mathbb{A}^1_e$ changes, the disjoint subvarieties Y_e cover an open subset V of X, so they are fibers of a proper fibration $V \to T/\mathbb{G}_m$. In particular, the singularities of $(Y_e, \Delta \cap Y_e)$ are no worse than singularities of (X, Δ), where Δ is the complement of the dense torus orbit in X.

3. In fact, each Y is isomorphic to $\mathbb{P}^{r-1}$, and $Y \cap \Delta = B_1 \cup B_2 \cup \cdots \cup B_n$ is the union of n hyperplanes in $\mathbb{P}^{r-1}$ in general position. The divisors B_i correspond to the n coordinate hyperplanes in $E = (\mathbb{A}^1)^n$.

4. By the Gelfand–MacPherson correspondence, the T-orbits of $\mathrm{Gr}^0(r, n)$ are in a natural bijection with PGL_r-orbits of n hyperplanes in $\mathbb{P}^{r-1}$ that are in general position, i.e. with isomorphism classes of labeled hyperplane arrangements $(\mathbb{P}^{r-1}, B_1 + \cdots + B_n)$. So, $U = U_{\mathrm{Gr}(r,n),\Delta(r,n)}$ is the moduli space of the general-position hyperplane arrangements.

Now look at any *stable* toric variety $X \to \mathrm{Gr}(r, n)$ of type $|\mathcal{Q}| = \Delta(r, n)$ and at a corresponding subvariety Y_e. Then the properties (1) and (2) above still hold. In particular, each Y is a generic section of X, and the singularities of $(Y, B_1 + \cdots + B_n = Y \cap \Delta)$ are no worse than the singularities of the pair (X, Δ). But by Lemma 4.4 the latter are semi log canonical. This implies that each pair $(Y, B_1 + \cdots + B_n)$ is stable in the sense of Definition 1.1.

Question 6.1. Can the moduli spaces $M_{Y,Q}$ in the case when Y is a partial flag variety, for example a variety of two-step flags, be interpreted as the moduli space of stable maps?

We also note that [7] provides an interpretation of the morphism

$$M_{\mathrm{Gr}(r,n),\Delta(r,n)} \to M_{\mathbb{P}(\Lambda^e E),\Delta(r,n)}$$

as a toric analogue of the extended Torelli map $\overline{M}_g \to \bar{A}_g$.

7. Higher Gromov–Witten theory

One of the exciting new frontiers for the moduli of stable pairs is the "higher-dimensional" Gromov–Witten theory, obtained by replacing the n-pointed stable curves $(X, B_1 + \cdots + B_n) \to Y$ by stable pairs with $\dim X \geq 2$. We list several questions in this direction.

Question 7.1. One way to define "higher" Gromov–Witten-invariants is to use evaluations at the intersections points $\bigcap_{j\in J} B_j$ with $|J| = \dim X$. Can a "generalized" quantum cohomology ring be defined using these evaluations? And is it a richer structure than simply an associative ring?

Question 7.2. Is there a more nontrivial definition, using evaluations at divisors rather than at points?

Question 7.3. Do the moduli spaces of weighted n-pointed curves, constructed by Hassett [27] lead to new ways to compute ordinary Gromov–Witten-invariants and descendants? When one varies the weights b_j and the moduli space $\overline{M}_{0,n}^{(b_j)}(\beta, Y)$ changes, is there a nice "wall-crossing" formula?

Question 7.4. The formula for the intersection products of ψ-classes on $\overline{M}_{0,n}$ is particularly easy (these are just the multinomial coefficients). What is the generalization of this formula for the compactified moduli space of hyperplane arrangements? Can it be obtained by using the toric degeneration of $\mathrm{Gr}(r, n)$ to a Gelfand–Tsetlin toric variety Z and thus degenerating $M_{\mathrm{Gr}(r,n),Q}$ to $M_{Z,Q}$?

References

[1] Abramovich, D., and Vistoli, A., Complete moduli for fibered surfaces. In *Recent progress in intersection theory* (Bologna, 1997), Trends Math., Birkhäuser Boston, Boston, MA, 2000, 1–31.

[2] Alexeev, V., Boundedness and K^2 for log surfaces. *Internat. J. Math.* **5** (6) (1994), 779–810.

[3] —, Log canonical singularities and complete moduli of stable pairs. arXiv alg-geom/9608013, 1996.

[4] —, Moduli spaces $M_{g,n}(W)$ for surfaces. In *Higher-dimensional complex varieties* (Trento, 1994), Walter de Gruyter, Berlin 1996, 1–22.

[5] —, On extra components in the functorial compactification of A_g. In *Moduli of abelian varieties* (Texel Island, 1999), Progr. Math. 195, Birkhäuser, Basel 2001, 1–9.

[6] —, Complete moduli in the presence of semiabelian group action. *Ann. of Math.* (2) **155** (3) (2002), 611–708.

[7] —, Compactified Jacobians and Torelli map. *Publ. Res. Inst. Math. Sci.* **40** (4) (2004), 1241–1265.

[8] Alexeev, V., Birkenhake, Ch., and Hulek, K., Degenerations of Prym varieties. *J. Reine Angew. Math.* **553** (2002), 73–116.

[9] Alexeev, V., and Brion, M., Stable reductive varieties. I. Affine varieties. *Invent. Math.* **157** (2) (2004), 227–274.

[10] —, Stable reductive varieties. II. Projective case. *Adv. Math.* **184** (2) (2004), 380–408.

[11] —, Moduli of affine schemes with reductive group action. *J. Algebraic Geom.* **14** (1) (2005), 83–117.

[12] —, Stable spherical varieties and their moduli. arXiv math.AG/0505673, 2005.

[13] Alexeev, V., and Knutson, A., Complete moduli spaces of branchvarieties. Preprint, 2006.

[14] Alexeev, V., and Mori, S., Bounding singular surfaces of general type. In *Algebra, arithmetic and geometry with applications* (West Lafayette, IN, 2000), Springer-Verlag, Berlin 2004, 143–174.

[15] Alexeev, V., and Nakamura, I., On Mumford's construction of degenerating abelian varieties. *Tôhoku Math. J.* (2) **51** (3) (1999) 399–420.

[16] Artin, M., Versal deformations and algebraic stacks. *Invent. Math.* **27** (1974), 165–189.

[17] Bravi, P., and Pezzini, G., Wonderful varieties of type *D*. *Represent. Theory* **9** (2005), 578–637.

[18] Brion, M., Sur l'image de l'application moment. In *Séminaire d'algèbre Paul Dubreil et Marie-Paule Malliavin* (Paris, 1986), Lecture Notes in Math. 1296, Springer-Verlag, Berlin 1987, 177–192.

[19] Deligne, P., and Mumford, D., The irreducibility of the space of curves of given genus. *Inst. Hautes Études Sci. Publ. Math.* **36** (1969), 75–109.

[20] Gelfand, I. M., Kapranov, M. M., and Zelevinsky, A. V., *Discriminants, resultants, and multidimensional determinants*. Math. Theory Appl., Birkhäuser, Boston, MA, 1994.

[21] Gwena, T., Degenerations of cubic threefolds and matroids. *Proc. Amer. Math. Soc.* **133** (5) (2005), 1317–1323.

[22] Hacking, P., Compact moduli of plane curves. *Duke Math. J.* **124** (2) (2004), 213–257.

[23] Hacking, P., Keel, S., and Tevelev, J., Compactification of the moduli space of hyperplane arrangements. arXiv math.AG/0509567, 2005.

[24] Haiman, M., and Sturmfels, B., Multigraded Hilbert schemes. *J. Algebraic Geom.* **13** (4) (2004), 725–769.

[25] Hassett, B., Stable log surfaces and limits of quartic plane curves. *Manuscripta Math.* **100** (4) (1999), 469–487.

[26] —, Stable limits of log surfaces and Cohen-Macaulay singularities. *J. Algebra* **242** (1) (2001), 225–235.

[27] —, Moduli spaces of weighted pointed stable curves. *Adv. Math.* **173** (2) (2003), 316–352.

[28] Hassett, B., and Kovács, S. J., Reflexive pull-backs and base extension. *J. Algebraic Geom.* **13** (2) (2004), 233–247.

[29] Kapranov, M. M., Chow quotients of Grassmannians. I. In *I. M. Gel'fand Seminar*, Adv. Soviet Math. 16, Amer. Math. Soc., Providence, RI, 1993, 29–110.

[30] Karu, K., Minimal models and boundedness of stable varieties. *J. Algebraic Geom.* **9** (1) (2000), 93–109.

[31] Keel, S., and Mori, S., Quotients by groupoids. *Ann. of Math.* (2) **145** (1) (1997), 193–213.

[32] Knudsen, F. F., The projectivity of the moduli space of stable curves. II, III. *Math. Scand.* **52** (2) (1983), 161–199, 200–212.

[33] Knudsen, F. F., and Mumford, D., The projectivity of the moduli space of stable curves. I. Preliminaries on "det" and "Div". *Math. Scand.* **39** (1) (1976), 19–55.

[34] Kollár, J., Projectivity of complete moduli. *J. Differential Geom.* **32** (1) (1990), 235–268.

[35] —, Adjunction and discrepancies. Flips and abundance for algebraic threefolds. *Astérisque* **211** (1992), 183–192.

[36] —, Quotient spaces modulo algebraic groups. *Ann. of Math.* (2) **145** (1) (1997), 33–79.

[37] —, Singularities of pairs. In *Algebraic geometry—Santa Cruz 1995*, Proc. Sympos. Pure Math. 62, Amer. Math. Soc., Providence, RI, 1997, 221–287.

[38] Kollár, J., and Mori, S., *Birational geometry of algebraic varieties*. Cambridge Tracts in Math. 134, Cambridge University Press, Cambridge 1998.

[39] Kollár, J., and Shepherd-Barron, N. I., Threefolds and deformations of surface singularities. *Invent. Math.* **91** (2) (1988), 299–338.

[40] Lafforgue, L., *Chirurgie des grassmanniennes*. CRM Monogr. Ser. 19, Amer. Math. Soc., Providence, RI, 2003.

[41] Luna, D., Variétés sphériques de type *A*. *Publ. Math. Inst. Hautes Études Sci.* **94** (2001), 161–226.

[42] Mumford, D., An analytic construction of degenerating abelian varieties over complete rings. *Compositio Math.* **24** (1972), 239–272.

[43] Namikawa, Y., A new compactification of the Siegel space and degeneration of Abelian varieties. I, II. *Math. Ann.* **221** (2–3) (1976), 97–141, 201–241.

[44] Oda, T., *Convex bodies and algebraic geometry*. Ergeb. Math. Grenzgeb. (3) 15, Springer-Verlag, Berlin 1988.

[45] Oxley, J. G., *Matroid theory*. Oxford Sci. Publ., The Clarendon Press/Oxford University Press, New York 1992.

[46] Peeva, I., and Stillman, M., Toric Hilbert schemes. *Duke Math. J.* **111** (3) (2002), 419–449.

[47] van Opstall, M. A., Moduli of products of curves. *Arch. Math.* (Basel) **84** (2) (2005), 148–154.

[48] Viehweg, E., *Quasi-projective moduli for polarized manifolds*. Ergeb. Math. Grenzgeb. 30, Springer-Verlag, Berlin 1995.

[49] Vologodsky, V., The locus of indeterminacy of the Prym map. *J. Reine Angew. Math.* **553** (2002), 117–124.

[50] Voronoi, G., Nouvelles applications des paramètres continus à la théorie des formes quadratique, I, II, III. *J. Reine Angew. Math.* **133** (1908), 97–178; **134** (1908), 198–287; **136** (1909), 67–181.

Department of Mathematics, University of Georgia, Athens, GA 30602, U.S.A.
E-mail: valery@math.uga.edu

Evaluation maps, slopes, and algebraicity criteria

Jean-Benoît Bost

Abstract. We discuss criteria for the algebraicity of a formal subscheme $\hat{V}$ in the completion $\hat{X}_P$ at some rational point P of an algebraic variety X over some field K. In particular we consider the case where K is a function field or a number field, and we discuss applications concerning the algebraicity of leaves of algebraic foliations, algebraic groups, absolute Tate cycles, and the rationality of germs of formal functions on a curve over a number field.

Mathematics Subject Classification (2000). 11G35, 11Jxx, 14G40, 14B20, 14F30, 14L15, 37F75.

Keywords. Arakelov geometry, algebraic foliations, algebraic groups, algebraicity and rationality criteria, capacity, Diophantine approximation and slopes.

1. Introduction

This article presents a survey of the *arithmetic algebraicity criteria* and their applications that have been developed in [12], [13], and in the subsequent joint work with Chambert-Loir [14].

The proofs of these criteria rely on what might be called the *method of slopes*, that is inspired by the classical techniques of auxiliary polynomials in Diophantine approximation, but is formulated in a geometric framework. In this approach, when investigating a projective algebraic variety X equipped with some ample line bundle L defined over a number field K, and some zero-dimensional subschemes Σ_i of X, the basic objects of interest are the *evaluation maps*

$$\eta_{D,i} : \Gamma(X, L^{\otimes D}) \longrightarrow \Gamma(\Sigma_i, L^{\otimes D})$$

which map global sections of $L^{\otimes D}$ to their restrictions to Σ_i. Typically, when X is the compactification of an algebraic group G, the Σ_i's may be some sets of multiples of some rational points of G, or some thickenings of such subsets. In the situation we shall deal with in this paper, the Σ_i's will be the successive infinitesimal neighbourhoods of some point P of $X(K)$ in a formal subscheme $\hat{V}$ of the formal completion $\hat{X}_P$.

The geometry of the Σ_i's in X turns out to be reflected by the injectivity properties of these evaluation maps – this is the contents of the so-called zero lemmas when X is some compactified algebraic group, and, in the setting of this paper, of the algebraicity criterion in Proposition 2.1 below. This geometry is finally related to the arithmetic properties of the data X, L, and Σ_i through the *slopes inequalities* satisfied by the

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

K-linear maps $\eta_{D,i}$. Indeed, after the choice of auxiliary data (such as integral models for X, L, and the Σ_i's, and hermitian metrics on $X(\mathbb{C})$ and $L_{\mathbb{C}}$), the source and range of $\eta_{D,i}$ appear as the underlying K-vector spaces attached to some hermitian vector bundles over $\mathrm{Spec}\,\mathcal{O}_K$. To them, elementary Arakelov geometry attaches arithmetic invariants, such as the *height* of $\eta_{D,i}$ and the *Arakelov degree* and the *slopes* of these hermitian vector bundles. The slope inequality for $\eta_{D,i}$ asserts that, when for instance $\eta_{D,i}$ is injective, the maximal slope of its source is bounded from above by the sum of its height and of the maximal slope of its range.

The approach to proving Diophantine statements by the consideration of evaluation maps and of the associated slope estimates has been introduced in a Bourbaki report [10], devoted to the work by Masser and Wüstholz on periods and minimal abelian subvarieties of abelian varieties over number fields [35]. The flexibility of this new geometric approach allowed the author to combine the original arguments in [35], phrased in terms of classical theta functions, with the "modern" theory of abelian schemes (including deep results due to Néron, Mumford, and Moret-Bailly), and to establish variants of the original work of Masser and Wüstholz where constants occuring in various estimates are explicitly bounded.

These effective versions of the "period theorem" of Masser and Wüstholz and of the consequent "isogeny estimates" have been recently improved by Viada ([43], [42]), by means of the same techniques. The combination of the method of slopes and of the modern theory of abelian schemes has also been used by Gaudron ([26]) to derive effective estimates on linear forms in logarithms on abelian varieties.

The results in the present article have been inspired by the work of D. and G. Chudnovsky [19], [18], and by the generalization of the results in [19] to abelian varieties by Graftieaux in [27] and [28], who also used the above combination of techniques. However, in the formulation and the proofs of the results discussed below, the flexibility of the method of slopes has not been exploited to derive Diophantine statements on abelian varieties involving *explicit* estimates, but instead to establish results valid in some *general geometric* setting by means of relatively non-technical arguments, at the expense of explicitness.

Another illustration of the flexibility of the method of slopes, in a spirit similar to this paper, is provided by the recent work by Gasbarri [25], who used this method to derive generalizations of transcendance theorems *à la* Schneider–Lang–Bombieri in general geometric situations and to clarify their relations with higher dimensional Nevanlinna theory.

In this article, we shall focus on these geometric aspects, instead of going into the details of the arguments of Arakelov geometry involved in proofs. For those, we refer to the original papers [12], [13], and [14] and to Chambert-Loir's Bourbaki report [16], which also discusses the link between slopes inequalities and more traditional techniques, such as Siegel's lemma and the interpolation determinants of Laurent [33].

To emphasize the geometric content of our approach, we shall first explain how the methods of auxiliary polynomials, in the guise of the study of the maps $\eta_{D,i}$ above, provides a simple algebraicity criterion for formal germs in an arbitrary projective

variety over a field K (section 2). Then, assuming that K is the function field $k(C)$ of some projective curve over some field k, we shall derive some geometric analogues of the results presented in the later sections. The proofs of these analogues will demonstrate how inequalities between slopes – of vector bundles over the curve C in this geometric setting – may be used to establish that, under suitable positivity conditions, the hypotheses of our previous algebraicity criterion are fulfilled (Section 3). Albeit technically simpler in the function field case, these arguments will give some insight into the proof of the arithmetic algebraicity criteria stated in Sections 6 and 7.

Besides, we shall illustrate these arithmetic criteria by applications to the algebraicity of leaves of *algebraic foliation*. Here, by an algebraic foliation over some base field K, we mean a smooth algebraic variety X over K, equipped with some sub-vector bundle F of the tangent bundle T_X, that is involutive (*i.e.*, whose sheaf of sections is closed under Lie bracket). When K is a field of characteristic $p > 0$, the sheaf of sections of T_X is equipped with the operation of p-th power, and it makes sense to require the sheaf of section of F to be closed under this operation. When K is a field of characteristic zero, for any point P in $X(K)$, one may consider the formal leaf of the foliation (X, F) through P, namely the unique smooth formal subscheme $\hat{V}$ of dimension rk F in the completion $\hat{X}_P$ whose formal tangent bundle coincides with the restriction of F.

When K is a number field with ring of integers $\mathcal{O}_K$, we may introduce some smooth model $\mathcal{X}$ of X over an open subscheme S of Spec $\mathcal{O}_K$ such that F extends to a sub-vector bundle $\mathcal{F}$ of $T_{\mathcal{X}/S}$, and consider the following condition, which we shall call the *Grothendieck–Katz condition*:

For almost every maximal ideal $\mathfrak{p}$ *in* Spec $\mathcal{O}_K$, *of residue characteristic* p, *the involutive subbundle* $\mathcal{F}_{\mathbb{F}_p}$ *of* $T_{\mathcal{X}_{\mathbb{F}_p}}$ *is stable under* p*-th power.*

It is easily seen to be satisfied when the foliation (X, F) is algebraically integrable[1]. The *generalized conjecture of Grothendieck–Katz* asserts that the converse holds. It was initially stated for linear differential systems by Katz [31] who attributes it to Grothendieck. The general formulation of the conjecture is due to Ekedahl, Shepherd-Barron, and Taylor [24]. Its investigation has been one of the main motivations behind the algebraicity results presented in this survey (see in particular, Section 6 and Theorems 6.1 and 6.2 below) which may also be considered as extensions of the earlier works of Chudnovsky ([18]) and André ([2], Chapter VIII, and [4], Section 5) on the original conjecture of Grothendieck–Katz.

For lack of space, we do not attempt to give any complete historical account of the origins of the algebraization techniques discussed below. Let us however indicate that these techniques – based on the consideration of maps sending global sections of ample line bundles to their restrictions to thickened points – goes back at least to the paper of Poincaré [37], where he presented an overview of his main results concerning abelian functions (see especially its Section II). Besides, the original proof of Chow's

[1] This means, by definition, that for any field extension Ω of K, the formal leaf through any point of $X(\Omega)$ of the algebraic foliation (X_Ω, F_Ω) is itself algebraic.

theorem relies on a criterion somewhat in the spirit of Proposition 2.1 below (see [17], Theorem IV). One might also refer to the article of Siegel [40] for historical comments on the proofs of algebraization statements in the framework of analytic geometry during the "pre-GAGA" era.

Algebraicity criteria in a geometric setting involving positivity conditions in the spirit of Theorem 3.1 below may also be deduced from classical results by Andreotti and Grauert on fields of meromorphic functions ([6], [5]) and by Hironaka, Matsumura, and Hartshorne on fields of formal meromorphic functions ([30], [29]); see [12], Section 3.3, and [9].

We refer to the monograph [7] for additional references concerning these techniques of formal geometry and their applications to extension and connectedness problems in projective geometry over a field. Many of these applications may be expected to have arithmetic counterparts which would extend the results in this article.

Conventions. The following notation and terminology are used throughout the article.

By an *algebraic scheme* over some field k, we mean a separated scheme of finite type over k. Integral algebraic schemes X over k will be called *algebraic varieties* over k.

Let G be an algebraic group over a field K of characteristic 0. Its Lie algebra Lie G is the fiber at the unit element $e \in G(K)$ of the tangent bundle T_G, and may be identified with the K-vector space of the left-invariant regular sections of T_G over G. The Lie bracket on Lie G is, by definition, the restriction of the Lie bracket on vector fields in $\Gamma(G, T_G)$. A Lie subalgebra $\boldsymbol{h}$, defined over K, of Lie G is called an *algebraic Lie subalgebra* when it is the Lie algebra of some algebraic subgroup H in G. When this holds, the subgroup H may be supposed connected, and then is unique and defined over any field of definition of G and $\boldsymbol{h}$.

If K is a number field, its ring of integers will be denoted $\mathcal{O}_K$. For any non-zero prime ideal $\boldsymbol{p}$ of $\mathcal{O}_K$, we let $N\boldsymbol{p} := |\mathcal{O}_K/\boldsymbol{p}|$ its norm, $K_{\boldsymbol{p}}$ (resp. $\mathcal{O}_{\boldsymbol{p}}$) the $\boldsymbol{p}$-adic completion of K (resp. of $\mathcal{O}_K$), and $|\ |_{\boldsymbol{p}}$ the $\boldsymbol{p}$-adic absolute value on $K_{\boldsymbol{p}}$ normalized in such a way that, for any uniformizing element ϖ of $\mathcal{O}_{\boldsymbol{p}}$, $|\varpi|_{\boldsymbol{p}} = N\boldsymbol{p}^{-1}$. We shall also denote K_v the completion of K at some place v (possibly archimedean).

2. Algebraic formal germs and auxiliary polynomials

Let X be an algebraic scheme over a field K, P a point of $X(K)$, $\hat{X}_P$ the formal completion of X at P, and $\hat{V} \hookrightarrow \hat{X}_P$ a smooth formal subscheme. Such a $\hat{V}$ will also be called a *smooth formal germ of subvariety* through P in X. For any non-negative integer i, we shall denote V_i the i-th infinitesimal neighborhood of P in $\hat{V}$. Thus,

$$V_0 = \{P\} \subset V_1 \subset V_2 \subset \cdots$$

and

$$\hat{V} = \varinjlim V_i.$$

We may consider the Zariski closure of $\hat{V}$ in X, namely, the smallest closed subscheme Z of X such that $\hat{Z}_P$ contain $\hat{V}$. Observe that it is a subvariety of X containing P: the ideal in $\mathcal{O}_{X,P}$ defining its germ at P is the intersection of $\mathcal{O}_{X,P}$ and of the ideal in its completion $\hat{\mathcal{O}}_{X,P} = \mathcal{O}_{\hat{X}_P}$ that defines $\hat{V}$, hence is prime. Moreover, since $\hat{Z}_P$ contains $\hat{V}$, the dimension of Z is at least equal to the dimension of $\hat{V}$.

The formal germ $\hat{V}$ is called *algebraic* when these two dimensions are equal. Indeed, using the compatibility properties of completion and normalization, one easily checks that this condition is equivalent to each of the following ones, which could have been used as alternative definitions:

(i) There exists a closed subvariety Z of X such that P belongs to $Z(K)$ and $\hat{V}$ is a branch of Z through P (*i.e.*, a component of the completion $\hat{Z}_P$).

(ii) There exist an integral algebraic scheme Y over K, a point 0 of $Y(K)$ and a K-morphism $f\colon Y \to X$ which maps 0 to P, such that the induced morphism on formal completions

$$\hat{f}_0\colon \hat{Y}_0 \longrightarrow \hat{X}_P$$

factorizes through $\hat{V} \hookrightarrow \hat{X}_P$ and defines a formal isomorphism from $\hat{Y}_0$ to $\hat{V}$.

Let us moreover assume that X *is projective* over K. We may choose an ample line bundle L on X, and introduce the following K-vector spaces and K-linear maps, for any non-negative integers D and i:

$$E_D := \Gamma(X, L^{\otimes D}),$$

$$\begin{aligned} \eta_D\colon E_D &\longrightarrow \Gamma(\hat{V}, L^{\otimes D}) \\ s &\longmapsto s_{|\hat{V}}, \end{aligned}$$

$$\begin{aligned} \eta_D^i\colon E_D &\longrightarrow \Gamma(V_i, L^{\otimes D}) \\ s &\longmapsto s_{|V_i}, \end{aligned}$$

and[2]

$$E_D^i := \{s \in E_D \mid s_{V_{i-1}} = 0\} = \ker \eta_D^{i-1}.$$

Observe that there is a canonical isomorphism $\Gamma(\hat{V}, L^{\otimes D}) \simeq \varprojlim_i \Gamma(V_i, L^{\otimes D})$, by means of which the map η_D gets identified with $\varprojlim_i \eta_D^i$.

The subspaces E_D^i define a decreasing filtration of E_D:

$$E_D = E_D^0 \supset E_D^1 \supset \cdots \supset E_D^i \supset E_D^{i+1} \supset \cdots.$$

Since the K-vector space E_D is finite dimensional, this filtration is stationary, and the very definition of Z as the Zariski closure of $\hat{V}$ shows that, if $\mathcal{I}_Z$ denotes its ideal sheaf in $\mathcal{O}_X$, we have

$$\bigcap_{i \geq 0} E_D^i = \ker \eta_D = \Gamma(X, \mathcal{I}_Z.L^{\otimes D}). \tag{2.1}$$

[2] In this definition, when $i = 0$, we let $V_{-1} = \emptyset$ and $\eta_D^{-1} = 0$.

Finally, if $T_{\hat{V}}$ denotes the tangent space of $\hat{V}$, then, for any non-negative integer i, the kernel of the restriction map from $\Gamma(V_i, L^{\otimes D})$ to $\Gamma(V_{i-1}, L^{\otimes D})$ may be identified with $S^i \check{T}_{\hat{V}} \otimes L_P^D$, and the restriction of the evaluation map η_D^i to E_D^i defines a K-linear map:

$$\gamma_D^i \colon E_D^i \longrightarrow S^i \check{T}_{\hat{V}} \otimes L_P^{\otimes D}.$$

Roughly speaking, it is the map which sends a section of $L^{\otimes D}$ vanishing up to order i at P along $\hat{V}$ to the $(i+1)$-th "Taylor coefficient" of its restriction to $\hat{V}$. By construction,

$$\ker \gamma_D^i = E_D^{i+1}. \tag{2.2}$$

The following proposition shows how the algebraicity of $\hat{V}$ and the asymptotic behaviour of the ranks of the subquotients E_D^i/E_D^{i+1} are related. It may be considered as a geometric version of the techniques of "auxiliary polynomials" used in the theory of Diophantine approximation: in our setting, the elements of the space $E_D := \Gamma(X, L^{\otimes D})$ play the role of auxiliary polynomials of degree D.

Proposition 2.1. *The following three conditions are equivalent:*

(i) *the formal germ $\hat{V}$ is algebraic;*

(ii) *there exists $c > 0$ such that, for any $(D, i) \in \mathbb{N}^2$ satisfying $i > cD$, the map γ_D^i vanishes;*

(iii) *the ratio*

$$\frac{\sum_{i\geq 0}(i/D)\mathrm{rk}\,(E_D^i/E_D^{i+1})}{\sum_{i\geq 0}\mathrm{rk}\,(E_D^i/E_D^{i+1})} \tag{2.3}$$

does not *admit the limit $+\infty$ when D goes to infinity.*

Condition (ii) may be also expressed by saying that, for every positive integer D the filtration $(E_D^i)_{i\geq 0}$ becomes stationary – or equivalently that η_D vanishes on E_D^i – when $i > cD$.

The implication (i) $\Rightarrow$ (ii) is a straightforward consequence of the basic theory of ample line bundles and their Seshadri constants (see for instance [34], Chapter 5, notably Proposition 5.1.9). The implication (ii) $\Rightarrow$ (iii) is clear. We sketch the proof of (ii) $\Rightarrow$ (i) below. The one of the implication (iii) $\Rightarrow$ (i) – which constitutes the algebraicity criterion we shall use in the sequel – is similar, but slightly more elaborate; see [13], Section 2.2, for more details.

Let us assume that condition (ii) holds, and let d denote the dimension of $\hat{V}$, which we may assume positive. Then, for any non-negative integers D and i, the quotient vector space $E_D^i/E_D^{i+1} = E_D^i/\ker \gamma_D^i \simeq \mathrm{im}\,\gamma_D^i$ has rank at most $\mathrm{rk}\,(S^i \check{T}_{\hat{V}} \otimes L_P^D) = \binom{d+i-1}{i}$ and vanishes if $i > cD$. This implies that

$$\mathrm{rk}\left(E_D/\bigcap_{i\geq 0} E_D^i\right) = \sum_{i\geq 0}\mathrm{rk}\,(E_D^i/E_D^{i+1}) \leq \sum_{i=0}^{[cD]}\binom{d+i-1}{i}.$$

Moreover the last sum is equivalent to $\frac{c^d}{d!}D^d$ when D goes to infinity.

Besides, according to (2.1),

$$E_D \Big/ \bigcap_{i\geq 0} E_D^i = \Gamma(X, L^{\otimes D}) / \Gamma(X, \mathfrak{I}_Z . L^{\otimes D}).$$

For D large enough, this space may be identified with $\Gamma(Z, L^{\otimes D})$ and its rank is equivalent to $\frac{\deg_L Z}{(\dim Z)!} D^{\dim Z}$ when D goes to infinity.

This shows that $\dim Z$ is at most d, and therefore is equal to d. This establishes condition (i), and completes the proof of (ii) $\Rightarrow$ (i).

3. An algebraicity criterion for smooth formal germs in varieties over function fields

Let C be a smooth projective, geometrically connected curve over some field k and let $K := k(C)$ be the associated function field.

Recall that, if E is a vector bundle of positive rank on C, its *slope* is defined as the quotient of its degree by its rank

$$\mu(E) := \frac{\deg E}{\operatorname{rk} E},$$

and its *maximal slope* $\mu_{\max}(E)$ is the maximum of the slopes $\mu(F)$ of sub-vector bundles of positive rank in E. Observe that, if L is any line bundle on C,

$$\mu_{\max}(E \otimes L) = \mu_{\max}(E) + \deg L.$$

Moreover, if E_1 and E_2 are vector bundles over C with E_2 of positive rank, and if there exists some (generically) injective morphism of vector bundles $\varphi\colon E_1 \to E_2$, then the following slope inequality holds:

$$\deg E_1 \leq \operatorname{rk} E_1 . \mu_{\max}(E_2). \tag{3.1}$$

Finally, recall that a vector bundle E over C is ample iff it has positive rank and there exists $c > 0$ such that, for any non-negative integer i,

$$\mu_{\max}(S^i \check{E}) \leq -c.i.$$

Let X be an algebraic scheme over K, P a point in $X(K)$, and $\hat{V} \subset \hat{X}_P$ a smooth formal germ of subvariety through P in X.

After possibly shrinking X, we may assume that it is quasi-projective and choose a quasi-projective model[3] $\pi : \mathcal{X} \longrightarrow C$ such that P extends to a section $\mathcal{P}$ of π.

[3] namely, a quasi-projective k-variety $\mathcal{X}$, equipped with a flat k-morphism $\pi : \mathcal{X} \to C$ and an isomorphism of its generic fiber $\mathcal{X}_K$ with X.

Theorem 3.1 ([13], Theorem 2.5). *With the above notation, assume that the following two conditions are satisfied:*

(i) *the formal subscheme $\hat{V}$ in $\hat{X}_P$ extends to a formal subscheme $\hat{\mathcal{V}}$ of $\hat{\mathcal{X}}_{\mathcal{P}}$ that is smooth over C;*

(ii) *the normal bundle $N_{\mathcal{P}}\hat{\mathcal{V}}$ of $\hat{\mathcal{V}}$ along $\mathcal{P}$ is ample.*

Then $\hat{V}$ is algebraic.

This algebraicity criterion may be seen as a "geometric model", concerning functions fields, of the arithmetic algebraicity criterion in Theorem 6.1 below, devoted to formal germs in varieties over number fields.

Our proof of Theorem 3.1 will rely on the implication (iii) $\Rightarrow$ (i) in Proposition 2.1. Indeed there exists a projective compactification of $\mathcal{X}$ to which the morphism π extends. Therefore we may assume that $\mathcal{X}$ is projective, and choose some ample line bundle $\mathcal{L}$ on $\mathcal{X}$. Let $L := \mathcal{L}_K$ be its restriction to X, and let E_D, E_D^i, η_D^i, and γ_D^i be as in the previous section. We are going to show that, when conditions (i) and (ii) in Theorem 3.1 are satisfied, the ratio (2.3) stays bounded when D goes to infinity.

To achieve this, let us consider the direct images $\mathcal{E}_D := \pi_*\mathcal{L}^{\otimes D}$ and $\pi_{|\mathcal{V}_i *}\mathcal{L}^{\otimes D}$, where $\mathcal{V}_i$ denotes the i-th infinitesimal neighbourhood of $\mathcal{P}$ in $\mathcal{V}$. These are torsion free coherent sheaves – or equivalently vector bundles – on C, which at the generic point Spec K of C coincide with the K-vector spaces E_D and $\Gamma(V_i, L^{\otimes D})$. Moreover, every restriction map $\eta_D^i \colon E_D \longrightarrow \Gamma(V_i, L^{\otimes D})$ extends to a morphism of vector bundles:

$$\begin{aligned} \overline{\eta}_D^i \colon \mathcal{E}_D &\longrightarrow \pi_{|\mathcal{V}_i *}\mathcal{L}^{\otimes D} \\ s &\longmapsto s_{|\mathcal{V}_i}. \end{aligned}$$

The filtration $(E_D^i)_{i\geq 0}$ of E_D also extends to the filtration of $\mathcal{E}_D$ by the sub-vector bundles $\mathcal{E}_D^i := \ker \overline{\eta}_D^{i-1}$. Finally, the kernel of the restriction map from $\pi_{|\mathcal{V}_i *}\mathcal{L}^{\otimes D}$ to $\pi_{|\mathcal{V}_{i-1} *}\mathcal{L}^{\otimes D}$ may be identified with $S^i(\check{N}_{\mathcal{P}}\hat{\mathcal{V}}) \otimes \mathcal{P}^*\mathcal{L}^{\otimes D}$ and the restriction of the evaluation map $\overline{\eta}_D^i$ to $\mathcal{E}_D^i$ defines a morphism of vector bundles $\overline{\gamma}_D^i \colon \mathcal{E}_D^i \to S^i(\check{N}_{\mathcal{P}}\hat{\mathcal{V}}) \otimes \mathcal{P}^*\mathcal{L}^{\otimes D}$, which coincides with γ_D^i at the generic point of C. The kernel of $\overline{\gamma}_D^i$ is $\mathcal{E}_D^{i+1}$ and therefore $\overline{\gamma}_D^i$ factorizes through a (generically) injective morphism of vector bundles $\tilde{\gamma}_D^i \colon \mathcal{E}_D^i/\mathcal{E}_D^{i+1} \to S^i(\check{N}_{\mathcal{P}}\hat{\mathcal{V}}) \otimes \mathcal{P}^*\mathcal{L}^{\otimes D}$.

The ampleness of $N_{\mathcal{P}}\hat{\mathcal{V}}$ and the slope inequality (3.1) applied to the morphisms $\tilde{\gamma}_D^i$ now show that, for some $c > 0$ independent of i and D, we have:

$$\deg(\mathcal{E}_D^i/\mathcal{E}_D^{i+1}) \leq \mathrm{rk}\,(E_D^i/E_D^{i+1})(-c.i + D.\deg \mathcal{P}^*\mathcal{L}).$$

Besides, since $\mathcal{L}$ is ample, the sheaf $\mathcal{E}_D$ is generated by its global sections for D large enough, and consequently:

$$\deg\Big(\mathcal{E}_D \Big/ \bigcap_{i\geq 0} \mathcal{E}_D^i\Big) \geq 0.$$

Moreover we may write:

$$\deg\left(\mathcal{E}_D\Big/\bigcap_{i\geq 0}\mathcal{E}_D^i\right)=\sum_{i\geq 0}\deg\left(\mathcal{E}_D^i/\mathcal{E}_D^{i+1}\right).$$

When D is large enough, the above three inequalities establish that

$$c.\sum_{i\geq 0}(i/D)\,\mathrm{rk}\,(E_D^i/E_D^{i+1})\leq \deg \mathcal{P}^*\mathcal{L}.\sum_{i\geq 0}\mathrm{rk}\,(E_D^i/E_D^{i+1}),$$

and finally that the ratio (2.3) is at most $(\deg \mathcal{P}^*\mathcal{L})/c$.

The following application of Theorem 3.1 illustrates how the algebraicity criteria contained in Proposition 2.1 and Theorem 3.1 lead to non-trivial geometric consequences in spite of the elementary nature of their proof (see also [9], [20], and [32] for applications of similar techniques to algebraic and rationally connected leaves of algebraic foliations).

Theorem 3.2 ([13], Theorem 2.6). *Let C be a smooth projective, geometrically connected curve over a field k of characteristic zero, $K := k(C)$ its function field, and $\pi : \mathcal{G} \to C$ a smooth group scheme over C. Let $G := \mathcal{G}_K$ be its generic fiber, and* Lie $\mathcal{G}$ *its Lie algebra*[4].

If the sub-vector bundle of Lie $\mathcal{G}$ *defined by some Lie subalgebra (over K) $\boldsymbol{h}$ of* Lie G *is ample, then it is an algebraic Lie subalgebra. More specifically, there exists a unipotent linear K-algebraic subgroup H in G such that $\boldsymbol{h} =$* Lie H.

In the classical analogy between function fields and number fields, this statement may be considered as a counterpart of Diophantine results concerning algebraic groups over number fields such as Theorem 6.3 below.

Observe also that, if G is a semi-abelian variety over K (hence does not admit any non-trivial unipotent algebraic subgroup), Theorem 3.2 asserts the semi-negativity of Lie $\mathcal{G}$. When $\mathcal{G}$ is an abelian scheme over C, this also follows from a classical curvature argument due to Griffiths.

Under the hypotheses of Theorem 3.2, to establish the existence of an algebraic subgroup H of G such that $\boldsymbol{h} =$ Lie H, one applies Theorem 3.1 in the situation where X is G, P is the unit element e of $G(K)$, and $\hat{V}$ is the "formal exponential" of the Lie subalgebra $\boldsymbol{h}$, namely, the formal subgroup of $\hat{G}_e$ such that $T_P\hat{V} = \boldsymbol{h}$. Its Zariski closure provides the sought for algebraic subgroup.

When the Lie bracket vanishes on $\boldsymbol{h}$, one may consider the vector group $\mathrm{Vect}(\boldsymbol{h})$ defined by $\boldsymbol{h}$ and the product $G' := \mathrm{Vect}(\boldsymbol{h})\times G$. The graph of the injection $\boldsymbol{h} \hookrightarrow$ Lie G is an algebraic Lie subalgebra of Lie G', and is the Lie algebra of the graph of an isomorphism $\mathrm{Vect}(\boldsymbol{h}) \simeq H \hookrightarrow G$.

[4]It is defined as the restriction along the zero section of π of the relative tangent bundle T_π. It is a vector bundle over C, equipped with a $\mathcal{O}_C$-bilinear Lie bracket, which coincides at the generic point Spec K of C with the Lie bracket of the Lie algebra Lie $G \simeq (\text{Lie } \mathcal{G})_K$ of the K-algebraic group G.

In general, one establishes that H is linear and unipotent by a similar argument, after having deduced that $\boldsymbol{h}$ is nilpotent from analyzing the compatibility of the Lie bracket on $\boldsymbol{h}$ with the Harder–Narasimhan filtration of the associated sub-vector bundle in Lie $\mathcal{G}$.

4. Sizes of formal subschemes over p-adic and global fields

This section and the next one are devoted to preliminaries needed for stating our arithmetic algebraicity criteria in Sections 6 and 7.

We first describe some constructions introduced in [12], Section 3, and further developed in [13], Section 4.1, and [14], Section 2. We refer to these papers for details and proofs.

Let k be a p-adic field (*i.e.*, a finite extension of $\mathbb{Q}_p$), $\mathcal{O}$ its subring of integers (*i.e.*, the integral closure of $\mathbb{Z}_p$ in k), $|\ |\colon k \to \mathbb{R}_+$ its absolute value, and $\mathbb{F}$ its residue field[5].

4.1. Groups of formal and analytic automorphisms. If $g := \sum_{I \in \mathbb{N}^d} a_I X^I$ is a formal power series in $k[[X_1, \ldots, X_d]]$ and if $r \in \mathbb{R}_+^*$, we define

$$\|g\|_r := \sup_I |a_I| r^{|I|} \in \mathbb{R}_+ \cup \{+\infty\}.$$

The "norm" $\|g\|_r$ is finite iff the series g is convergent and bounded on the open ball of radius r in $\overline{k}^d$.

Let $\hat{\mathbb{A}}_k^d$ be the formal completion at the origin of the d-dimensional affine space over k. Its group $\mathrm{Aut}\,\hat{\mathbb{A}}_k^d$ of automorphisms may be identified with the space of d-tuples $f = (f_i)_{1 \leq i \leq d}$ of formal series $f_i \in k[[x_1, \ldots, x_d]]$ such that $f(0) = 0$ and $Df(0) := \left(\frac{\partial f_i}{\partial x_j}(0)\right)_{1 \leq i,j \leq d}$ belongs to $\mathrm{GL}_n(k)$.

We shall consider the following subgroups of $\mathrm{Aut}\,\hat{\mathbb{A}}_k^d$:

- the subgroup G_{for} formed by the formal automorphisms f such that $Df(0)$ belongs to $\mathrm{GL}_n(\mathcal{O})$;
- the subgroup G_ω formed by the elements $f := (f_i)_{1 \leq i \leq d}$ of G_{for} such that the series f_i have positive radii of convergence;
- for any $r \in \mathbb{R}_+^*$, the subgroup $G_\omega(r)$ of G_ω formed by the elements $f := (f_i)_{1 \leq i \leq d}$ of G_{for} such that the series f_i satisfy the bounds $\|f_i\|_r \leq r$.

The group $G_\omega(r)$ may be seen as the group of analytic automorphisms, defined over k and preserving the origin, of the open d-dimensional ball of radius r. Moreover

[5] Actually we might assume more generally that k is any field equipped with a complete non-Archimedean absolute value $|\ |\colon k \to \mathbb{R}_+$ and let $\mathcal{O} := \{t \in k \mid |t| \leq 1\}$ be its valuation ring.

we have:

$$r' > r > 0 \Longrightarrow G_\omega(r') \subset G_\omega(r) \quad \text{and} \quad \bigcup_{r>0} G_\omega(r) = G_\omega.$$

4.2. The size $S_{\mathcal{X}}(\hat{V})$ of a formal germ $\hat{V}$. The filtration $(G_\omega(r))_{r>0}$ of the group G_ω may be used to attach a number $S_{\mathcal{X}}(\hat{V})$ in $[0, 1]$ to any smooth formal germ $\hat{V}$ in an algebraic variety X over k, depending on the choice of some model $\mathcal{X}$ of X over $\mathcal{O}$. This number shall provide some quantitative measure of the analyticity of $\hat{V}$.

Let $\hat{V}$ be a formal subscheme of $\hat{\mathbb{A}}^d_k$. For any φ in Aut $\hat{\mathbb{A}}^d_k$, we may consider its inverse image $\varphi^*(\hat{V})$, which also is a formal subscheme of $\hat{\mathbb{A}}^d_k$. Observe that $\hat{V}$ is a smooth formal scheme of dimension v iff there exists φ in Aut $\hat{\mathbb{A}}^d_k$ such that $\varphi^*(\hat{V})$ is the formal subscheme $\hat{\mathbb{A}}^v_k \times \{0\}$ of $\hat{\mathbb{A}}^d_k$. Moreover, when this holds, φ may be chosen in G_{for}.

Similarly, the formal germ $\hat{V}$ is *analytic and smooth* – namely, it is the formal scheme attached to some germ at 0 of smooth analytic subspace of dimension v of the d-dimensional affine space over k – iff there exists φ in G_ω such that $\varphi^*(\hat{V})$ is the formal subscheme $\hat{\mathbb{A}}^v_k \times \{0\}$ of $\hat{\mathbb{A}}^d_k$.

These observations lead us to introduce the *size of a smooth formal subscheme* $\hat{V}$ of dimension v of $\hat{\mathbb{A}}^d_k$, defined as the supremum $S(\hat{V})$ in $[0, 1]$ of the real numbers $r \in]0, 1]$ for which there exists φ in $G_\omega(r)$ such that $\varphi^*(\hat{V})$ is the formal subscheme $\hat{\mathbb{A}}^v_k \times \{0\}$ of $\hat{\mathbb{A}}^d_k$.

More generally, if $\mathcal{X}$ is an $\mathcal{O}$-scheme of finite type equipped with a section $\mathcal{P} \in \mathcal{X}(\mathcal{O})$ and if $\hat{V}$ is a smooth formal subscheme of the formal completion $\hat{X}_P$ of $X := \mathcal{X}_k$ at $P := \mathcal{P}_k$, then the *size* $S_{\mathcal{X}}(\hat{V})$ *of* $\hat{V}$ *with respect to the model* $\mathcal{X}$ of X will be defined as the size of $i(\hat{V})$, where $i\colon U \hookrightarrow \mathbb{A}^d_{\mathcal{O}}$ is an embedding of some open neighbourhood U in $\mathcal{X}$ of the section $\mathcal{P}$ into an affine space of large enough dimension d, which additionally maps $\mathcal{P}$ to the origin $0 \in \mathbb{A}^d_{\mathcal{O}}(\mathcal{O})$.

This definition is independent of the choices of U, d, and i, and extends the previous one. Actually it satisfies the following invariance properties:

I1. If $\mathcal{X}$ is a subscheme of a scheme $\mathcal{X}'$ over $\mathcal{O}$, then $S_{\mathcal{X}'}(\hat{V}) = S_{\mathcal{X}}(\hat{V})$.

I2. If $\mathcal{X}$, X, $\mathcal{P}$, $\hat{V}$ and $\mathcal{X}'$, X', $\mathcal{P}'$, $\hat{V}'$ are as above, and if there exists an $\mathcal{O}$-morphism $\phi\colon \mathcal{X} \to \mathcal{X}'$ mapping $\mathcal{P}$ to $\mathcal{P}'$, étale along $\mathcal{P}$, such that the formal isomorphism $\hat{\phi}_k\colon \hat{X}_P \xrightarrow{\sim} \hat{X}'_{P'}$ maps isomorphically $\hat{V}$ onto $\hat{V}'$, then $S_{\mathcal{X}'}(\hat{V}') = S_{\mathcal{X}}(\hat{V})$.

Besides, the size $S_{\mathcal{X}}(\hat{V})$ is invariant by extension of the p-adic base field k (*cf.* [14]).

Finally observe that, with the same notation as above, the size $S_{\mathcal{X}}(\hat{V})$ is positive iff $\hat{V}$ is analytic. Moreover, if $\hat{V}$ extends to a formal subscheme $\hat{\mathcal{V}}$ of the formal completion of $\mathcal{X}$ along $\mathcal{P}$ which is smooth along $\mathcal{P}$, then $S_{\mathcal{X}}(\hat{V}) = 1$.

4.3. Size of formal leaves of algebraic foliations. It is possible to establish lower bounds on the sizes of formal germs of solutions of algebraic ordinary differential equations. These bounds will allow us to apply our arithmetic algebraicity criteria below to the solutions of algebraic differential equations – or more generally, to leaves of algebraic foliations – defined over number fields.

Proposition 4.1. *Let $\mathcal{X}$ be a smooth scheme over* Spec $\mathcal{O}$*, $\mathcal{P}$ a section in $\mathcal{X}(\mathcal{O})$, and $\mathcal{F}$ a sub-vector bundle of rank f in $T_{\mathcal{X}/\mathcal{O}}$. Let us assume that the subbundle $F := \mathcal{F}_k$ of the tangent bundle T_X of the smooth k-variety $X := \mathcal{X}_k$ is involutive, and let $\hat{V}$ be the formal germ of leave of this involutive bundle through $P := \mathcal{P}_k$.*

1) *The size of $\hat{V}$ with respect to $\mathcal{X}$ satisfies the lower bound:*

$$S_{\mathcal{X}}(\hat{V}) \geq |\pi| := |p|^{\frac{1}{p-1}}. \tag{4.1}$$

2) *If moreover k is absolutely unramified and if the reduction $\mathcal{F}_{\mathbb{F}} \hookrightarrow T_{\mathcal{X}_{\mathbb{F}}}$ of $\mathcal{F}$ to the closed fiber $\mathcal{X}_{\mathbb{F}}$ of $\mathcal{X}$ is closed under p-th power, then*

$$S_{\mathcal{X}}(\hat{V}) \geq |p|^{\frac{1}{p(p-1)}}. \tag{4.2}$$

This is proved in [12], Proposition 3.9, and [13], Proposition 4.1, by first reducing the construction of $\hat{V}$ to the one of the formal flow $\psi_{(t_1,\dots,t_f)}$ of suitably chosen commuting sections $v_1, \dots, v_f$ of F. Then one studies the analyticity properties of this flow by expanding the map $\psi^*_{(t_1,\dots,t_f)}$ (defined by $\psi_{(t_1,\dots,t_f)}$ acting on functions) *à la* Cauchy:

$$\psi^*_{(t_1,\dots,t_f)} = \exp\Big(\sum_{1\leq i\leq f} t_i.v_i\Big) := \sum_{(i_1,\dots,i_f)\in\mathbb{N}^f} \frac{t_1^{i_1}\dots t_f^{i_f}}{i_1!\dots i_f!} D_1^{i_1}\circ\dots\circ D_f^{i_f},$$

where $D_1, \dots, D_f$ denote the derivations of the sheaf of regular functions on X defined by $v_1, \dots, v_f$.

4.4. A-germs. Consider an algebraic variety X over some number field K, P a point in $X(K)$, and $\hat{V}$ a smooth formal subscheme in $\hat{X}_P$.

Let N be a positive integer and $(\mathcal{X}, \mathcal{P})$ a model of (X, P) over $\mathcal{O}_K[1/N]$. For any maximal ideal $\boldsymbol{p}$ in $\mathcal{O}_K$ not dividing N, by base change we get a smooth formal germ $\hat{V}_{K_{\boldsymbol{p}}}$ through $P_{K_{\boldsymbol{p}}}$ in the algebraic variety $X_{K_{\boldsymbol{p}}}$ over the p-adic field $K_{\boldsymbol{p}}$, and a model $(\mathcal{X}_{\mathcal{O}_{\boldsymbol{p}}}, \mathcal{P}_{\mathcal{O}_{\boldsymbol{p}}})$ over $\mathcal{O}_{\boldsymbol{p}}$ of the pair $(X_{K_{\boldsymbol{p}}}, P_{K_{\boldsymbol{p}}})$. Consequently, for any such $\boldsymbol{p}$, the size $S_{\mathcal{X}_{\mathcal{O}_{\boldsymbol{p}}}}(\hat{V}_{K_{\boldsymbol{p}}})$ is a well-defined element in $[0, 1]$.

We shall say that the formal germ $\hat{V}$ in X is *A-analytic*, or is an *A-germ*, when the following two conditions are satisfied:

1. *for any place v of K, the formal germ $\hat{V}_{K_v}$ is K_v-analytic*[6];

[6] or equivalently, if for any maximal ideal $\boldsymbol{p}$ of $\mathcal{O}_K$, the formal germ $\hat{V}_{K_{\boldsymbol{p}}}$ is $K_{\boldsymbol{p}}$-analytic in $X_{K_{\boldsymbol{p}}}$, and for any complex embedding $\sigma : K \hookrightarrow \mathbb{C}$, $\hat{V}_\sigma$ is $\mathbb{C}$-analytic in X_σ.

2. *the infinite product* $\prod_{\mathfrak{p}\nmid N} S_{\mathcal{X}_{\mathcal{O}_\mathfrak{p}}}(\hat{V}_{K_\mathfrak{p}})$ *is positive*, or equivalently,

$$\sum_{\mathfrak{p}\nmid N} \log S_{\mathcal{X}_{\mathcal{O}_\mathfrak{p}}}(\hat{V}_{K_\mathfrak{p}})^{-1} < +\infty.$$

This pair of conditions does not depend on the choices of the integer N and the model $(\mathcal{X}, \mathcal{P})$. Moreover, it is invariant under extension of the base field[7].

Recall that, if X is a variety over a number field K and P is some smooth point in $X(K)$, then the G-functions at the point P of X are the elements f in the completion $\mathcal{O}_{\hat{X}_P}$ of $\mathcal{O}_{X,P}$ defined by similar conditions: the analyticity at every place of K, and the positivity of the infinite product as in Condition 2 above, where $S_{\mathcal{X}_{\mathcal{O}_\mathfrak{p}}}(\hat{V}_{K_\mathfrak{p}})$ is replaced by $\min(1, R_\mathfrak{p})$, $R_\mathfrak{p}$ denoting the $\mathfrak{p}$-adic radius of convergence of f expressed in some fixed system of local coordinates on X at P (see for instance [2] and [22] for details and references).

It is straightforward that, if the graph $\mathrm{Gr}\, f$ of some f in $\mathcal{O}_{\hat{X}_P}$ – this graph is a smooth formal germ through $P' := (P, f(P))$ in $X' := X \times \mathbb{A}^1$ – is A-analytic, then f is a G-function. Let us emphasize that the converse does *not* hold[8].

Observe also that an algebraic smooth formal germ is always A-analytic. Even more, if $\hat{V}$ is an algebraic smooth formal germ in $\hat{X}_P$, where as above X denotes an algebraic variety over some number field K and P a point in $X(K)$, and if N is a positive integer and $(\mathcal{X}, \mathcal{P})$ some model of (X, P) over $\mathrm{Spec}\, \mathcal{O}_K[1/N]$, then almost all the sizes $S_{\mathcal{X}_{\mathcal{O}_\mathfrak{p}}}(\hat{V}_{K_\mathfrak{p}})$ are equal to one. Indeed, after "shrinking" $\mathrm{Spec}\, \mathcal{O}_K[1/N]$ to $\mathrm{Spec}\, \mathcal{O}_K[1/N']$, with N' a suitable multiple of N, the formal scheme $\hat{V}$ extends to a formal subscheme $\hat{\mathcal{V}}$ of the formal completion of $\mathcal{X}$ along $\mathcal{P}$ which is smooth along $\mathcal{P}$. (In substance, this observation goes back to the last memoir of Eisenstein [23], which may be considered as the starting point of the arithmetic theory of differential equations.)

Finally, observe that Proposition 4.1 admits the following straightforward consequence:

Corollary 4.2. *If X is a smooth algebraic variety over some number field K and if F is an involutive subbundle of T_X that satisfies the Grothendieck–Katz condition* (*see Introduction*)*, then the formal germ of leave of the so-defined algebraic foliation through any point P in $X(K)$ is A-analytic.*

[7] Namely, with the above notation, for any finite degree extension L of K, the formal germ $\hat{V}_L$ through P_L in the algebraic variety X_L over the number field L is A-analytic iff $\hat{V}$ is.

[8] For instance, the series $\log(1+x) := \sum_{n=1}^{+\infty} x^n/n \in \mathbb{Q}[[x]]$ defines a G-function at the point 0 in $\mathbb{A}^1_\mathbb{Q}$. However, its graph coincides with the transpose of the graph of the series $\exp y - 1 := \sum_{n=1}^{+\infty} y^n/n!$, which is not a G-function, and consequently is not an A-germ.

5. Condition L and canonical semi-norms

5.1. Consistent sequences of norms. Let k be a local field, X a projective scheme over k, and L a line bundle over X.

We may consider the following natural constructions of sequences of norms on the spaces of sections $\Gamma(X, L^{\otimes n})$:

1. When k is a p-adic field, with ring of integer $\mathcal{O}$, we may choose a pair $(\mathcal{X}, \mathcal{L})$, where $\mathcal{X}$ is a projective flat model of X over $\mathcal{O}$, and $\mathcal{L}$ a line bundle over $\mathcal{X}$ extending L. Then, for any integer n, the $\mathcal{O}$-module $\Gamma(\mathcal{X}, \mathcal{L}^{\otimes n})$ is (torsion-)free of finite rank and may be identified with an $\mathcal{O}$-lattice in the k-vector space $\Gamma(X, L^{\otimes n})$, and consequently defines a norm on the latter – namely, the norm $\| \, . \, \|_n$ such that a section $s \in \Gamma(X, L^{\otimes n})$ satisfies $\|s\|_n \leq 1$ iff s extends to a section of $\mathcal{L}^{\otimes n}$ over $\mathcal{X}$.

2. When $k = \mathbb{C}$ and X is reduced, we may consider any continuous norm $\| \, . \, \|_L$ on the $\mathbb{C}$-analytic line bundle L_{an} defined by L on the compact and reduced complex analytic space $X(\mathbb{C})$. Then, for any integer n, the space of algebraic regular sections $\Gamma(X, L^{\otimes n})$ may be identified with a subspace of the space of continuous sections of $L_{\mathrm{an}}^{\otimes n}$ over $X(\mathbb{C})$. Thus it is endowed with the restriction of the L^∞-norm, defined by:

$$\|s\|_{\mathrm{L}^\infty, n} := \sup_{x \in X(\mathbb{C})} \|s(x)\|_{L^{\otimes n}} \quad \text{for any } s \in \Gamma(X, L^{\otimes n}), \tag{5.1}$$

where $\| \, . \, \|_{L^{\otimes n}}$ denotes the continuous norm on $L_{\mathrm{an}}^{\otimes n}$ deduced from $\| \, . \, \|_L$ by taking the n-th tensor power.

This construction admits a variant where, instead of the sup-norms (5.1), one considers the L^p-norms defined by using some "Lebesgue measure" (*cf.* [12], 4.1.3, and [38], Théorème 3.10).

3. When $k = \mathbb{R}$ and X is reduced, the previous constructions define complex norms on the complex vector spaces

$$\Gamma(X, L^{\otimes n}) \otimes_{\mathbb{R}} \mathbb{C} \simeq \Gamma(X_{\mathbb{C}}, L_{\mathbb{C}}^{\otimes n})$$

and, by restriction, real norms on the real vector spaces $\Gamma(X, L^{\otimes n})$.

For any given k, X, and L as above, we shall say that two sequences $(\| \, . \, \|_n)_{n \in \mathbb{N}}$ and $(\| \, . \, \|'_n)_{n \in \mathbb{N}}$ of norms on the finite k-dimensional vector spaces $(\Gamma(X, L^{\otimes n}))_{n \in \mathbb{N}}$ are *equivalent* when, for some positive constant C and any positive integer n,

$$C^{-n} \, \| \, . \, \|'_n \leq \| \, . \, \|_n \leq C^n \, \| \, . \, \|'_n \, .$$

One easily checks that the previous constructions provide sequences of norms $(\| \, . \, \|_n)_{n \in \mathbb{N}}$ on the spaces $(\Gamma(X, L^{\otimes n}))_{n \in \mathbb{N}}$ which are all equivalent. A sequence of norms on these spaces equivalent to one (or, equivalently, to any) of the sequences thus constructed will be called *consistent*. This notion immediately extends to sequences $(\| \, . \, \|_n)_{n \geq n_0}$ of norms on the spaces $\Gamma(X, L^{\otimes n})$ defined for n large enough.

When the line bundle L is ample, consistent sequences of norms are provided by additional constructions. Indeed we have:

Proposition 5.1. *Let k be a local field, X a projective scheme over k, and L an* ample *line bundle over X. Let moreover Y be a closed subscheme of X, and assume X and Y reduced when k is archimedean.*

For any consistent sequence of norms $(\|\,.\,\|_n)_{n\in\mathbb{N}}$ on $(\Gamma(X, L^{\otimes n}))_{n\in\mathbb{N}}$, the quotient norms $(\|\,.\,\|'_n)_{n\in\mathbb{N}}$ on the spaces $(\Gamma(Y, L_{|Y}^{\otimes n}))_{n\geq n_0}$, deduced from the norms $\|\,.\,\|_n$ by means of the restriction maps $\Gamma(X, L^{\otimes n}) \longrightarrow \Gamma(Y, L_{|Y}^{\otimes n})$ – which are surjective for $n \geq n_0$ large enough since L is ample – *constitute a consistent sequence.*

When k is archimedean, this is proved in [13], Appendix, by introducing a positive metric on L, as a consequence of Grauert's finiteness theorem for pseudo-convex domains applied to the unit disk bundle of $\check{L}$ (see also [38]). When k is a p-adic field with ring of integers $\mathcal{O}$, Proposition 5.1 follows from the basic properties of ample line bundles over projective $\mathcal{O}$-schemes.

Let E be a finite dimensional vector space over the local field k, equipped with some norm, supposed to be euclidean or hermitian in the archimedean case. This norm induces similar norms on the tensor powers $E^{\otimes n}$, $n \in \mathbb{N}$, hence – by taking the quotient norms – on the symmetric powers $S^n E$. If X is the projective space $\mathbb{P}(E) := \operatorname{Proj} \operatorname{Sym}^{\cdot}(E)$ and L the line bundle $\mathcal{O}(1)$, then the canonical isomorphisms $S^n E \simeq \Gamma(X, L^{\otimes n})$ allow one to see these norms as a sequence of norms on $(\Gamma(X, L^{\otimes n}))_{n\in\mathbb{N}}$. One easily checks that this sequence is consistent[9].

For any closed subvariety Y in $\mathbb{P}(E)$ and any $n \in \mathbb{N}$, we may consider the commutative diagram of k-linear maps:

$$\begin{array}{ccccc} S^n E & \xrightarrow{\sim} & S^n \Gamma(\mathbb{P}(E), \mathcal{O}(1)) & \xrightarrow{\sim} & \Gamma(\mathbb{P}(E), \mathcal{O}(n)) \\ & & \downarrow & & \downarrow{\scriptstyle \alpha_n} \\ & & S^n \Gamma(Y, \mathcal{O}(1)) & \xrightarrow{\beta_n} & \Gamma(Y, \mathcal{O}(n)) \end{array}$$

where the vertical maps are the obvious restriction morphisms. The maps α_n, and consequently β_n, are surjective if n is large enough.

Together with Proposition 5.1, these observations yield the following corollary:

Corollary 5.2. *Let k, E and Y a closed subscheme[10] of $\mathbb{P}(E)$ be as above. Let us choose a norm on E (resp. on $\Gamma(Y, \mathcal{O}(1))$) and let us equip $S^n E$ (resp. $S^n \Gamma(Y, \mathcal{O}(1))$) with the induced norm, for any $n \in \mathbb{N}$.*

Then the sequence of quotient norms on $\Gamma(Y, \mathcal{O}(n))$ defined, when n is large enough, by means of the surjective morphisms $\alpha_n \colon S^n E \longrightarrow \Gamma(Y, \mathcal{O}(n))$ (resp. by $\beta_n \colon S^n \Gamma(Y, \mathcal{O}(1)) \longrightarrow \Gamma(Y, \mathcal{O}(n))$) is consistent.

5.2. Conditions L and $\mathbf{L}_v$. Let k be a local field, and X a projective integral scheme over k, equipped with an ample line bundle L. Moreover let $\hat{V} \hookrightarrow \hat{X}_P$ be a smooth

[9] This is straightforward in the p-adic case. When k is archimedean, this follows for instance from [15], Lemma 4.3.6.

[10] reduced if k is archimedean.

formal germ in X through a point $P \in X(k)$, and consider its tangent space $T_P\hat{V}$, the fiber L_P of L at P, and the evaluation maps

$$\gamma_D^i : \Gamma(X, \mathcal{I}_{V_{i-1}} \otimes L^{\otimes D}) \longrightarrow S^i \check{T}_{\hat{V}} \otimes L_P^{\otimes D} \tag{5.2}$$

introduced in Section 2.

Let us choose a consistent sequence of norms $(\| \, . \, \|_n)_{n\in\mathbb{N}}$ on the k-vector spaces $(\Gamma(X, L^{\otimes n}))_{n\in\mathbb{N}}$, and arbitrary norms $\| \, . \, \|_{T_P\hat{V}}$ on $T_P\hat{V}$ and $\| \, . \, \|_{L_P}$ on L_P. Then we may consider the operator norms $\|\gamma_D^i\|$ of the maps (5.2) and their logarithms $\log \|\gamma_D^i\|$ in $[-\infty, +\infty[$.

We shall say that $\hat{V}$ *satisfies condition* L when

$$\lim_{i/D\to+\infty} \frac{1}{i} \log \|\gamma_D^i\| = -\infty. \tag{5.3}$$

Clearly this condition does not depend on the above choices of norms. It is also invariant by extension of the local field k, and is easily seen not to depend on the choice of the ample line bundle L. Proposition 5.1 also implies that it is invariant under "reembedding" of X into some larger projective variety.

Moreover condition L is birationally invariant in the following sense: if $X' \dashrightarrow X$ is a birational map between projective varieties over k that define an isomorphism $f : U' \xrightarrow{\sim} U$ between non-empty open subvarieties in X' and X, and if P belongs to $U(k)$, then a smooth formal germ $\hat{V}$ through P in X satisfies L iff the smooth formal germ $f^*\hat{V}$ through $P' := f_{|U'}^{-1}(P)$ in X' does. (See [13], 3.2, when k is archimedean and $\dim \hat{V} = 1$; the general case is similar.)

As a consequence, condition L makes sense for a smooth formal germ $\hat{V} \hookrightarrow \hat{X}_P$ through a k-rational point in a general algebraic variety X over k – namely, if U is any quasi-projective open neighbourhood of P in X and if $\tilde{X}$ is a projective completion of U, we shall say that $\hat{V}$ satisfies L when $\hat{V}$ seen as a formal subscheme of $\tilde{X}$ satisfies it. Again, this condition is invariant under extension of k and reembedding of X.

Finally, if K is a number field and v a place of K, we shall say that a smooth formal subscheme $\hat{V} \hookrightarrow \hat{X}_P$ through a rational point P in a variety X over K satisfies *condition* L_v when the formal subscheme $\hat{V}_{K_v}$ through P in X_{K_v}, deduced from $\hat{V}$ by extension of scalars from K to the completion K_v of K at v, satisfies condition L over the local field K_v.

5.3. Condition L over $\mathbb{C}$ and Liouville complex manifolds. A connected complex manifold M is said to satisfy the *Liouville property*, or to be a *Liouville complex manifold,* when every bounded plurisubharmonic function on M is constant. In particular, the connected Riemann surfaces satisfying the Liouville property are precisely the ones which are "parabolic" in the sense of Myrberg, or equivalently, have "null-boundary" in the sense of R. Nevanlinna.

The following observations are straightforward consequences of the basic properties of plurisubharmonic functions and algebraic varieties:

1. Let $\pi : M \longrightarrow N$ be a surjective analytic map between connected complex manifolds. If M is Liouville, then N is Liouville. Conversely, when π has smooth connected fibers, if N and the fibers of π are Liouville, then M also is Liouville.
2. The complement of any closed pluripolar subset (for instance, a lower dimensional analytic subset) in a Liouville complex manifold is again Liouville.
3. Any compact connected complex manifold is Liouville.
4. The manifold of complex points of any smooth connected complex algebraic variety is Liouville.
5. Any connected complex Lie group is a Liouville complex manifold.

Over archimedean local fields, the property L may be checked in various significant cases by means of the following criterion:

Proposition 5.3 ([12], Proposition 4.12)). *Let X be a complex algebraic variety, P a point in $X(\mathbb{C})$, and $\hat{V} \hookrightarrow \hat{X}_P$ a smooth formal germ through P.*

Let us assume that there exist a connected complex manifold M, a point O in M, and a $\mathbb{C}$-analytic map $\varphi : M \longrightarrow X(\mathbb{C})$ sending O to P that induces an isomorphism of formal germs

$$\hat{\varphi}_O : \hat{M}_O \xrightarrow{\sim} \hat{V}. \tag{5.4}$$

If furthermore M is Liouville, then $\hat{V}$ satisfies L.

5.4. Germs of analytic curves in algebraic varieties over local fields and canonical semi-norms. In this paragraph, we return to the notation of the beginning of 5.2, and we assume that the smooth formal germ $\hat{V}$ is *one-dimensional* and *k-analytic*. Then, by means of the evaluation maps γ_D^i (see (5.2)) and their operator norms, as in the definition of condition L by (5.3), we may define some *canonical semi-norm* $\| \, . \, \|_{P,\hat{V}}^{\mathrm{can}}$ on the k-line $T_P\hat{V}$ as follows. We consider

$$\rho := \limsup_{i/D \to +\infty} \frac{1}{i} \log \left\| \gamma_D^i \right\|.$$

A straightforward application of Cauchy's inequalities shows that it belongs to $[-\infty, +\infty[$, and therefore by setting

$$\| \, . \, \|_{P,\hat{V}}^{\mathrm{can}} := e^{\rho} \| \, . \, \|_{T_P\hat{V}},$$

we define a semi-norm on $T_P\hat{V}$.

For a given projective variety X containing $\hat{V}$, one easily checks that it depends neither on the auxiliary choices of norms, nor on the ample line bundle L. Actually, like condition L, the canonical semi-norm $\| \, . \, \|_{P,\hat{V}}^{\mathrm{can}}$ is invariant under "reembedding" of X in some larger projective variety, and by birational isomorphisms which are isomorphisms in some neighbourhood of P ([13], 3.2–3.3, and [14]). Consequently,

the canonical semi-norm on $T_P\hat{V}$ may be defined for any smooth analytic germ of curve $\hat{V} \hookrightarrow \hat{X}_P$ through a rational point in an algebraic scheme over k.

In the p-adic case, Cauchy's inequalities lead actually to the following upper bound on the canonical semi-norm in terms of the size relative to some model:

Lemma 5.4. *Let k be a p-adic field, $\mathcal{O}$ its ring of integers, and $\mathcal{X}$ a separated scheme of finite type over $\mathcal{O}$ equipped with a section P. Let $\hat{V}$ be a smooth formal subscheme of the formal completion $\hat{X}_{P_k}$ of $X := \mathcal{X}_k$ at P_k. If $\hat{V}$ is one-dimensional and analytic, and if $\|\,.\,\|^{\mathcal{X}}_{T_P\hat{V}}$ denotes the p-adic norm on the k-line $T_P\hat{V}$ defined by the integral model $\mathcal{X}$*[11]*, then we have:*

$$\|\,.\,\|^{\mathrm{can}}_{P,\hat{V}} \leq S_{\mathcal{X}}(\hat{V})^{-1}.\,\|\,.\,\|^{\mathcal{X}}_{T_P\hat{V}}\,.$$

Finally observe that the construction of $\|\,.\,\|^{\mathrm{can}}_{P,\hat{V}}$ is compatible with (finite degree) extensions of the local field k.

5.5. Canonical semi-norms and capacity. Recall that, if M is a Riemann surface, O a point of M, and Ω an open neighbourhood of O that is relatively compact in M and has a non-empty sufficiently regular boundary[12], we may consider the *Green function of O in Ω*, namely the continuous function $g_{O,\Omega}\colon \overline{\Omega} \setminus \{O\} \to \mathbb{R}_+$ which vanishes on the boundary of Ω, is harmonic on $\Omega \setminus \{O\}$, and possesses a logarithmic singularity at O. In other words, if z denotes some holomorphic coordinates on some open neighbourhood U of O, we have

$$g_{O,\Omega} = \log|z - z(O)|^{-1} + h \quad \text{on } U \setminus \{O\},$$

where h is a harmonic function on U. From the value of h at O, one defines the *capacitary norm* $\|\,.\,\|^{\mathrm{cap}}_{O,\Omega}$ on the complex line $T_O M = \mathbb{C}\frac{\partial}{\partial z}_{|O}$ by

$$\|\frac{\partial}{\partial z}_{|O}\|^{\mathrm{cap}}_{O,\Omega} := e^{-h(O)} = \lim_{Q \to O} \frac{e^{-g_{O\Omega}(Q)}}{|z(Q) - z(O)|}. \tag{5.5}$$

Proposition 5.5 ([13], Proposition 3.6). *With the above notation, if $f\colon \Omega \to X$ is a holomorphic map with value in some complex algebraic variety X, and if C denotes some germ of smooth analytic curve through $P := f(O)$ in X such that f maps the germ of Ω at O to C, then, for any v in $T_O M$,*

$$\|Df(O)v\|^{\mathrm{can}}_{P,C} \leq \|v\|^{\mathrm{cap}}_{O,\Omega}.$$

[11] By definition, the unit disk in $\check{T}_P\hat{V}$ equipped with the norm dual to $\|\,.\,\|^{\mathcal{X}}_{T_P\hat{V}}$ is the $\mathcal{O}$-lattice image of the composite map $P^*\Omega^1_{\mathcal{X}/\mathcal{O}} \to \Omega^1_{X/k|P_k} \twoheadrightarrow \check{T}_P\hat{V}$.

[12] say, a domain with differentiable non-empty boundary, or in other terms, the interior of some connected 2-dimensional submanifold with non-empty boundary. See also [11], especially A.8, for weaker conditions.

The construction of the Green function $g_{P,\Omega}$ admits analogues over p-adic curves, developed in particular by Rumely [39] and Thuillier [41] (see also [14] for a more "algebraic" approach relying on formal geometry). This Green function makes sense for instance when M is a smooth projective geometrically connected algebraic curve over some p-adic field k, O is some point in $M(k)$, and Ω is defined as the complement of some non-empty affinoid subspace of X that does not contain P. It has a logarithmic singularity at O and the equation (5.5) still makes sense and defines the capacitary norm as a p-adic norm on the k-line $T_O M$. Moreover Proposition 5.5 still holds with k instead of $\mathbb{C}$ as a base field, and "rigid analytic" instead of "holomorphic" (*cf.* [14], Sections 6 and 7).

6. An algebraicity criterion for smooth formal germs in varieties over number fields

6.1. An algebraization theorem. The following theorem provides sufficient conditions of algebraicity for a formal subscheme of the formal completion $\hat{X}_P$ of some algebraic variety X over a number field K at a rational point $P \in X(K)$.

Theorem 6.1. *Let X be an algebraic variety over a number field K, P a point in $X(K)$, and $\hat{V}$ a smooth formal subscheme of the completion $\hat{X}_P$ of X at P.*

If $\hat{V}$ is A-analytic and satisfies condition L_v for some place v of K, then $\hat{V}$ is algebraic.

This theorem is proved by using the algebraicity criterion (iii) $\Rightarrow$ (i) in Proposition 2.1. Validity of condition (iii) is derived by means of slope inequalities, involving now heights of K-linear maps and arithmetic slopes attached to hermitian vector bundles over $\operatorname{Spec} \mathcal{O}_K$, in the spirit of the proof of Theorem 3.1. See [12], Section 4 when the place v is archimedean; the general case is similar.

6.2. Algebraic leaves of algebraic foliations over number fields. Combined with Corollary 4.2 and Proposition 5.3, Theorem 6.1 admits the following consequence:

Theorem 6.2 ([12], Theorem 2.1). *Let X be a smooth algebraic variety over a number field K equipped with an involutive subbundle F, and let P be point in $X(K)$. If* (i) *the algebraic foliation (X, F) satisfies the Grothendieck–Katz condition, and* (ii) *for some field embedding $\sigma_0 \colon K \hookrightarrow \mathbb{C}$, the analytic leaf through P of the complex analytic foliation $(X(\mathbb{C}), F_{\mathbb{C}})$ is Liouville, then the leaf of (X, F) through P is algebraic.*

6.3. Algebraic Lie subalgebras of algebraic groups over number fields. Let G be an algebraic group over a number field K. For any sufficiently divisible integer N, there exists a model $\mathcal{G}$ of G, *i.e.*, a smooth quasi-projective group scheme over $S := \operatorname{Spec} \mathcal{O}_K[1/N]$ whose generic fiber $\mathcal{G}_K$ coincides with G. The restriction to the zero-section of $\mathcal{G}$ of the relative tangent bundle $T_{\mathcal{G}/S}$ defines the Lie algebra $\operatorname{Lie} \mathcal{G}$ of $\mathcal{G}$:

it is a finitely generated projective module and a Lie algebra over $\mathcal{O}_K[1/N]$, and the K-Lie algebra $(\mathrm{Lie}\,\mathcal{G})_K$ is canonically isomorphic to $\mathrm{Lie}\,G$.

Moreover, for every maximal ideal $\boldsymbol{p}$ of $\mathcal{O}_K[1/N]$ with residue field $\mathbb{F}_{\boldsymbol{p}}$ of characteristic p, the $\mathbb{F}_{\boldsymbol{p}}$-Lie algebra $(\mathrm{Lie}\,\mathcal{G})_{\mathbb{F}_{\boldsymbol{p}}}$ is canonically isomorphic to the Lie algebra of the smooth algebraic group $\mathcal{G}_{\mathbb{F}_{\boldsymbol{p}}}$ over the finite field $\mathbb{F}_{\boldsymbol{p}}$, and is therefore endowed with a p-th power map, given by the restriction of the p-th power map on global sections of $T_{\mathcal{G}_{\mathbb{F}_{\boldsymbol{p}}}}$ to the left-invariant ones.

For translation invariant foliations on G, Theorem 6.2 takes the following form which proves a conjecture of Ekedahl, Shepherd-Barron, and Taylor ([24]):

Theorem 6.3 ([12], Theorem 2.3). *For any Lie subalgebra* $\boldsymbol{h}$ *of* $\mathrm{Lie}\,G$ (*defined over* K)*, the following two conditions are equivalent:*

(i) *For almost every maximal ideal* $\boldsymbol{p}$ *of* $\mathcal{O}_K[1/N]$*, the* $\mathbb{F}_{\boldsymbol{p}}$*-Lie subalgebra* $(\boldsymbol{h} \cap \mathrm{Lie}\,\mathcal{G})_{\mathbb{F}_{\boldsymbol{p}}}$ *of* $\mathrm{Lie}\,\mathcal{G}_{\mathbb{F}_{\boldsymbol{p}}}$ *is closed under* p*-th powers.*

(ii) $\boldsymbol{h}$ *is an algebraic Lie subalgebra of* $\mathrm{Lie}\,G$.

6.4. Ogus conjecture on absolute Tate cycles in abelian varieties. Theorem 6.3 may be extended to the case where the field K is any extension of finite type of $\mathbb{Q}$, in the spirit of the original formulation of the Grothendieck–Katz conjecture [31]. In this section, we discuss some consequence of this generalization.

Let K be a field of characteristic zero, extension of finite type of $\mathbb{Q}$, and let X be a proper and smooth scheme over K. We can find models of X and K which are smooth over $\mathrm{Spec}\,\mathbb{Z}$ – namely an integral affine scheme $S = \mathrm{Spec}\,R$ smooth over $\mathrm{Spec}\,\mathbb{Z}$ such that K is the function field $\kappa(S)$ of S, and a proper and smooth scheme $\mathcal{X}$ over S such that $X = \mathcal{X}_K$. After possibly shrinking S, we can also assume that the Hodge cohomology groups $H^q(\mathcal{X}, \Omega^p_{\mathcal{X}/S})$ are flat R-modules. This implies that the formation of the (relative) Hodge and de Rham cohomology groups of $\mathcal{X}$ is compatible with any base change $S' \to S$ and the degeneracy of the "Hodge to de Rham" spectral sequence.

Let k be a perfect field of characteristic $p > 0$, W its ring of Witt vectors, $\overline{\sigma}$ a point in $S(k)$, and σ a lift of $\overline{\sigma}$ in $S(W)$. The crystalline cohomology groups $H^i_{\mathrm{cris}}(\mathcal{X}_{\overline{\sigma}}/W)$ are W-modules attached to the k-scheme $\mathcal{X}_{\overline{\sigma}}$. By functoriality, the absolute Frobenius endomorphism of $\mathcal{X}_{\overline{\sigma}}$ induces a Frobenius-linear endomorphism Φ of the W-module $H^i_{\mathrm{cris}}(\mathcal{X}_{\overline{\sigma}}/W)$. Besides, the comparison theorem of Berthelot ([8]) provides a canonical isomorphism from this W-module onto the de Rham cohomology $H^i_{\mathrm{dR}}(\mathcal{X}_\sigma/W) \simeq H^i_{\mathrm{dR}}(\mathcal{X}/R) \otimes_\sigma W$. Consequently, Φ defines a semi-linear endomorphism Φ_σ of the W-module $H^i_{\mathrm{dR}}(\mathcal{X}/R) \otimes_\sigma W$.

Let r be an integer in $\{0, \ldots, \dim X\}$. Following the terminology of Ogus in [36], a class ξ in the algebraic de Rham cohomology group $H^{2r}_{\mathrm{dR}}(\mathcal{X}/S)$ over R is said to be *absolutely Tate* iff, for any p, k, and σ as above[13], the equality

$$\Phi_\sigma(\xi) = p^r \xi$$

[13] Actually, one might consider only some "limited" classes of fields k and points $\overline{\sigma}$ in $\mathcal{X}(k)$ – for instance, closed points, or geometric generic points of the fibres of $S \to \mathrm{Spec}\,\mathbb{Z}$ – which still define the same condition.

holds in $H^{2r}_{\mathrm{dR}}(\mathcal{X}/S)\otimes_\sigma W$. More generally a class ξ in $H^{2r}_{\mathrm{dR}}(X/K)$ is said to be absolutely Tate if, after possibly replacing S by some non-empty affine open subscheme, it is absolutely Tate in $H^{2r}_{\mathrm{dR}}(\mathcal{X}/S)$.

For instance, any algebraic class, namely any class in the image of the "cycle class" map $Z^r(X)_{\mathbb{Q}} \longrightarrow H^{2r}_{\mathrm{dR}}(X/K)$, is absolutely Tate. Ogus ([36], Section 2) conjectured that the converse holds, that is:

O(X, r): *every absolutely Tate class in* $H^{2r}_{\mathrm{dR}}(X/K)$ *is algebraic.*

As a consequence of the aforementioned generalization[14] of Theorem 6.3 to algebraic groups over K, we can prove:

Theorem 6.4. *For any field extension K of finite type of $\mathbb{Q}$, the conjecture* O(X, r) *holds when X is an abelian variety over K and $r = 1$.*

See also [3], Section 7.4, for a discussion of an essentially equivalent theorem, which characterizes the K-linear maps between the first de Rham cohomology groups of abelian varieties over a number field K induced by morphisms of K-varieties, up to isogeny, as the ones compatible with (almost all) the crystalline Frobenius maps[15].

The derivation of Theorem 6.4 relies on the identification of the Lie algebra of the universal vector extension of the dual abelian variety $\hat{X}$ with the de Rham cohomology group $H^1_{\mathrm{dR}}(X/K)$, and on the fact that the p-th power map on the reduction of this Lie algebra at some closed point $\boldsymbol{p}$ in S of residue characteristic p coincides with the reduction modulo p of the crystalline Frobenius at $\boldsymbol{p}$.

7. An algebraicity criterion for smooth formal curves in varieties over number fields

7.1. Normed and semi-normed lines over number fields. We define a *normed line*

$$\bar{L} := (L_K, (\|\,.\,\|_{\boldsymbol{p}}), (\|\,.\,\|_\sigma))$$

over a number field K as the data of a rank one K-vector space L_K, of a family $(\|\,.\,\|_{\boldsymbol{p}})$ of $\boldsymbol{p}$-adic norms on the $K_{\boldsymbol{p}}$-lines $L_K\otimes_K K_{\boldsymbol{p}}$ indexed by the non-zero prime ideals $\boldsymbol{p}$ of $\mathcal{O}_K$, and of a family $(\|\,.\,\|_\sigma)$ of hermitian norms on the complex lines $L_K\otimes_{K,\sigma}\mathbb{C}$, indexed by the fields embeddings $\sigma\colon K\hookrightarrow\mathbb{C}$. Moreover the family $(\|\,.\,\|_\sigma)$ is required to be stable under complex conjugation[16].

[14] This generalization concerns an algebraic group G over K, and a smooth group scheme $\mathcal{G}$ extending it over S; in condition (i) in Theorem 6.3, one now requires the p-closure condition to hold for every closed point $\boldsymbol{p}$ in some non-empty open subscheme of S over which $\boldsymbol{h}$ extends as a subvector bundle of Lie $\mathcal{G}$. Its proof relies on the fact that a complex manifold is Liouville when it may be fibered over a complex algebraic variety with fibers some complex Lie groups.

[15] The author became aware of Ogus conjecture in the above formulation while reading a preliminary version of Y. André's beautiful survey on motives [3], and realized that, when K is a number field, X an abelian variety, and $r = 1$, it would be a consequence of Theorem 6.3. Actually, in [36], Section 2, Ogus formulates more general conjectures, stated in terms of the conjugate filtration on $H^*_{\mathrm{dR}}(X/K)$ and its reductions, and asks explicitly about the validity of O(X, r) only when K is a number field.

We shall say that a normed K-line is *summable* if for some (or equivalently, for any) non-zero element l of L_K, the family of real numbers $(\log \|l\|_{\boldsymbol{p}})_{\boldsymbol{p}}$ is summable. Then we may define its *Arakelov degree* as the real number

$$\widehat{\deg}\, \bar{L} := \sum_{\boldsymbol{p}} \log \|l\|_{\boldsymbol{p}}^{-1} + \sum_{\sigma} \log \|l\|_{\sigma}^{-1}. \tag{7.1}$$

Indeed, by the product formula, the right-hand side of (7.1) does not depend on the choice of l.

Observe that hermitian line bundles over $\operatorname{Spec} \mathcal{O}_K$, as usually defined in Arakelov geometry, provide examples of normed lines over K: if $\bar{\mathcal{L}} = (\mathcal{L}, (\|\,.\,\|_{\sigma})_{\sigma\colon K \hookrightarrow \mathbb{C}})$ is such an hermitian line bundle – so $\mathcal{L}$ is a projective $\mathcal{O}_K$-module of rank 1, and $(\|\,.\,\|_{\sigma})_{\sigma\colon K \hookrightarrow \mathbb{C}})$ is a family, invariant under complex conjugation, of norms on the complex lines $\mathcal{L}_{\sigma} := \mathcal{L} \otimes_{\sigma\colon \mathcal{O}_K \to \mathbb{C}} \mathbb{C}$ – the corresponding normed K-line is $\mathcal{L}_K$ equipped with the p-adic norms defined by the $\mathcal{O}_{\boldsymbol{p}}$-lattices $\mathcal{L} \otimes_{\mathcal{O}_K} \mathcal{O}_{\boldsymbol{p}}$ in $\mathcal{L} \otimes_{\mathcal{O}_K} K_{\boldsymbol{p}} \simeq L \otimes_K K_{\boldsymbol{p}}$ and with the hermitian norms $(\|\,.\,\|_{\sigma})$. The so-defined normed lines are summable, and their Arakelov degree, as defined by (7.1), coincide with the usual Arakelov degree of hermitian line bundles.

It is convenient to extend the definitions of normed lines and Arakelov degree as follows: we shall define a *semi-normed K-line* $\bar{L}$ as a K-vector space L_K of rank one, equipped with families of *semi-norms* $(\|\,.\,\|_{\boldsymbol{p}})$ and $(\|\,.\,\|_{\sigma})$, where the latter is assumed to be stable under complex conjugation. In other words, we allow some of the $\|\,.\,\|_{\boldsymbol{p}}$ or $\|\,.\,\|_{\sigma}$ to vanish.

We shall say that the Arakelov degree of a semi-normed K-line $\bar{L}$ is *well-defined* if, for some (or equivalently, for any), non-zero element l of L_K, the family of real numbers $(\log^+ \|l\|_{\boldsymbol{p}})_{\boldsymbol{p}}$ is summable. Then we may again define its Arakelov degree by means of (7.1), where we follow the usual convention $\log 0^{-1} = +\infty$. It is an element of $]-\infty, +\infty]$.

7.2. Arithmetic positivity and algebraicity of A-germs of curves. The following algebraicity criterion is a refined version of Theorem 6.1, concerning formal germs of curves:

Theorem 7.1 ([14]). *Let X be an algebraic variety over a number field K, P a point in $X(K)$, and $\hat{V}$ a smooth formal subscheme of dimension* 1 *in the completion $\hat{X}_P$ of X at P.*

Assume that the following two conditions are satisfied:

(i) *$\hat{V}$ is A-analytic;*

(ii) *the semi-normed K-line*

$$\overline{T}_P \hat{V}^{\mathrm{can}} := (T_P \hat{V}, (\|\,.\,\|^{\mathrm{can}}_{P, \hat{V}_{K_{\boldsymbol{p}}}}), (\|\,.\,\|^{\mathrm{can}}_{P, \hat{V}_{\sigma}})),$$

[16] The data of these families of norms is equivalent to the data of a family $(\|\,.\,\|_v)_v$, indexed by the set of all places v of K, of v-adic norms on the rank one vector spaces $L_v := L_K \otimes_K K_v$ over the v-adic completions K_v of K.

defined by endowing $T_P\hat{V}$ with its canonical semi-norm at every place of K, satisfies

$$\widehat{\deg}\,\overline{T}_P\hat{V}^{\mathrm{can}} > 0. \tag{7.2}$$

Then $\hat{V}$ is algebraic.

Observe that, as a consequence of Lemma 5.4, the Arakelov degree of $\overline{T}_P\hat{V}^{\mathrm{can}}$ is well defined in $]-\infty,+\infty]$ when Condition (i) is satisfied. Moreover it takes the value $+\infty$ if there exists some place v of K such that Condition L_v is satisfied.

7.3. A rationality criterion for formal germs of functions on algebraic curves over number fields. Let K be a number field, $\mathcal{C}$ a regular projective arithmetic surface over $\operatorname{Spec}\mathcal{O}_K$ whose generic fiber $C := \mathcal{X}_K$ is geometrically connected, P a point in $X(K)$, and $\mathcal{P}$ in $\mathcal{X}(\mathcal{O}_K)$ extending P.

Let F be a finite set of closed points in $\operatorname{Spec}\mathcal{O}_K$ and, for any $\boldsymbol{p}$ in F, let $\Omega_{\boldsymbol{p}}$ be the complement in the rigid curve $X_{K_{\boldsymbol{p}}}$ of some affinoid not containing $P_{K_{\boldsymbol{p}}}$. Moreover, for any embedding $\sigma\colon K \to \mathbb{C}$, let Ω_σ be an open neighbourhood of P_σ in the Riemann surface $C_\sigma(\mathbb{C})$, which for simplicity we suppose to be domains with differentiable non-empty boundaries. We shall assume that the data of the Ω_σ's are compatible with complex conjugation, namely, that for any embedding σ, $\Omega_{\bar{\sigma}}$ is the complex conjugate of Ω_σ.

Let finally $\overline{T}_P^{\mathrm{cap}}C := (T_PC, (\|\,.\,\|_{\boldsymbol{p}}), (\|\,.\,\|_\sigma))$ be the semi-normed line over K defined by the tangent line T_C of C at P, equipped with the $\boldsymbol{p}$-adic norm $\|\,.\,\|_{\boldsymbol{p}} := \|\,.\,\|^{\mathrm{cap}}_{P_{K_{\boldsymbol{p}}},\Omega_{\boldsymbol{p}}}$ if $\boldsymbol{p}$ belongs to F, and otherwise with the $\boldsymbol{p}$-adic norm deduced from the integral structure on $N_{\mathcal{P}}\mathcal{X}$ (through the isomorphism $T_PC\otimes_K K_{\boldsymbol{p}} \simeq N_{\mathcal{P}}\mathcal{X}\otimes_{\mathcal{O}_K} K_{\boldsymbol{p}}$). Its Arakelov degree is clearly defined.

The following theorem extends the classical rationality criteria of Borel–Dwork and Polya–Bertrandias (*cf.* [21] and [1], Chapter 5); one recovers them in the special case $C = \mathbb{P}^1_K$.

Theorem 7.2 ([14]). *With the above notation, let $\varphi \in \hat{\mathcal{O}}_{C,P}$ be a formal germ of function on C at P, and assume that the following conditions are satisfied:*

(i) *for any $\boldsymbol{p}$ in F (resp. any embedding $\sigma\colon k \hookrightarrow \mathbb{C}$), after the base change $K \hookrightarrow K_{\boldsymbol{p}}$ (resp. σ), φ extends to a rigid meromorphic function on $\Omega_{\boldsymbol{p}}$ (resp. to a meromorphic function on Ω_σ);*

(ii) *the formal function φ extends to a formal rational function on the completion $\hat{\mathcal{X}}_{\mathcal{P}}$ over* $\operatorname{Spec}\mathcal{O}_K \setminus F$*;*

(iii) $\widehat{\deg}\,\overline{T}_P^{\mathrm{cap}}C > 0$.

Then φ is rational, i.e.*, is an element of the local ring $\mathcal{O}_{X,P} \subset \hat{\mathcal{O}}_{X,P}$.*

To establish Theorem 7.2, we consider the graph of φ. It is a formal germ of curve through the point $(P,\varphi(P))$ in the algebraic variety $X := C\times\mathbb{P}^1_K$ over K. From

Theorem 7.1, Proposition 5.5, and its p-adic analogue, we derive that this graph is algebraic. Finally, we show that it is the germ of graph of some rational function on C by applying a generalization of the connectedness theorems in [11], Section 4.

References

[1] Amice, Y., *Les nombres p-adiques*. Presses Universitaires de France, Paris 1975.

[2] André, Y., *G-functions and geometry*. Friedr. Vieweg & Sohn, Braunschweig 1989.

[3] André, Y., *Une introduction aux motifs. Motifs purs, motifs mixtes, périodes*. Panor. Synthèses 17, Société Mathématique de France, Paris 2004.

[4] André, Y., Sur la conjecture des p-courbures de Grothendieck-Katz et un problème de Dwork. In *Geometric aspects of Dwork theory* (ed. by A. Adolphson et al.), Vol. I, Walter de Gruyter, Berlin 2004, 55–112.

[5] Andreotti, A., Théorèmes de dépendance algébrique sur les espaces complexes pseudo-concaves. *Bull. Soc. Math. France* **91** (1963), 1–38.

[6] Andreotti, A., and Grauert, H., Algebraische Körper von automorphen Funktionen. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II* **1961** (1961), 39–48.

[7] Bădescu, L., *Projective geometry and formal geometry*. IMPAN Monogr. Mat. (N.S.) 65, Birkhäuser Verlag, Basel 2004.

[8] Berthelot, P., *Cohomologie cristalline des schémas de caractéristique $p > 0$*. Lecture Notes in Math. 407, Springer-Verlag, Berlin 1974.

[9] Bogomolov, F., and McQuillan, M. L., Rational curves on foliated varieties. Preprint, I.H.E.S., 2001.

[10] Bost, J.-B., Périodes et isogénies des variétés abéliennes sur les corps de nombres (d'après D. Masser et G. Wüstholz). Séminaire Bourbaki 1994/95, Exp. no. 795. *Astérisque* **237** (1996), 115–161.

[11] Bost, J.-B., Potential theory and Lefschetz theorems for arithmetic surfaces. *Ann. Sci. École Norm. Sup.* **32** (1999), 241–312.

[12] Bost, J.-B., Algebraic leaves of algebraic foliations over number fields. *Inst. Hautes Études Sci. Publ. Math.* **93** (2001), 161–221.

[13] Bost, J.-B., Germs of analytic varieties in algebraic varieties: canonical metrics and arithmetic algebraization theorems. In *Geometric aspects of Dwork theory* (ed. by A. Adolphson et al.), Vol. I, Walter de Gruyter, Berlin 2004, 371–418

[14] Bost, J.-B., and Chambert-Loir, A., Analytic curves in algebraic varieties over number fields. Preliminary version, 2005.

[15] Bost, J.-B., Gillet, H., and Soulé, C., Heights of projective varieties and positive Green forms. *J. Amer. Math. Soc.* **7** (4) (1994), 903–1027, .

[16] Chambert-Loir, A., Théorèmes d'algébricité en géométrie diophantienne. Séminaire Bourbaki, Vol. 2000/2001, Exposé 886; *Astérisque* **282** (2002), 175–209.

[17] Chow, W.-L., On compact complex analytic varieties. *Amer. J. Math.* **71** (1949), 893–914.

[18] Chudnovsky, D. V., and Chudnovsky, G. V., Applications of Padé approximations to the Grothendieck conjecture on linear differential equations. In *Number theory* (New York, 1983–84), Lecture Notes in Math. 1135, Springer-Verlag, Berlin 1985, 52–100.

[19] Chudnovsky, D. V., and Chudnovsky, G. V., Padé approximations and Diophantine geometry. *Proc. Nat. Acad. Sci. U.S.A.* **82** (8) (1985), 2212–2216.

[20] Druel, S., Caractérisation de l'espace projectif. *Manuscr. Math.* **115** (1) (2004), 19–30.

[21] Dwork, B., On the rationality of the zeta function of an algebraic variety. *Amer. J. Math.* **82** (1960), 631–648.

[22] Dwork, B., Gerotto, G., and Sullivan, F. J., *An introduction to G-functions*. Princeton University Press, Princeton, NJ, 1994.

[23] Eisenstein, G., Eine allgemeine Eigenschaft der Reihen-Entwicklungen aller algebraischen Funktionen. *Bericht der Königl. Preuss. Akademie der Wissenschaften zu Berlin*, 1852, 441–443.

[24] Ekedahl, T., Shepherd-Barron, N. I., and Taylor, R., A conjecture on the existence of compact leaves of algebraic foliations. Available at http://www.dpmms.cam.ac.uk/~nisb /fruit/WWW/foliation.dvi, 1999.

[25] Gasbarri, C., Analytic subvarieties with many rational points. Preprint, Università di Roma "Tor Vergata", 2005.

[26] Gaudron, E., Formes linéaires de logarithmes effectives sur les variétés abéliennes (2004). Preprint, Université de Grenoble 1, 2004.

[27] Graftieaux, P., Formal groups and the isogeny theorem. *Duke Math. J.* **106** (1) (2001), 81–121.

[28] Graftieaux, P., Formal subgroups of abelian varieties. *Invent. Math.* **145** (1) (2001), 1–17.

[29] Hartshorne, R., Cohomological dimension of algebraic varieties. *Ann. of Math.* (2) **88** (1968), 403–450.

[30] Hironaka, H., and Matsumura, H., Formal functions and formal embeddings. *J. Math. Soc. Japan* **20** (1968), 52–82.

[31] Katz, N. M., Algebraic solutions of differential equations (p-curvature and the Hodge filtration). *Invent. Math.* **18** (1972), 1–118.

[32] Kebekus, S., Solà Conde, L., and Toma, M., Rationally connected foliations after Bogomolov and McQuillan. *J. Algebraic Geom.*, to appear.

[33] Laurent, M., Sur quelques résultats récents de transcendance. In *Journées arithmétiques* (Luminy, 1989); *Astérisque* **198–200** (1991), 209–230.

[34] Lazarsfeld, R., *Positivity in algebraic geometry. I. Classical setting: line bundles and linear series.* Ergeb. Math. Grenzgeb. 48, Springer-Verlag, Berlin 2004.

[35] Masser, D., and Wüstholz, G., Periods and minimal abelian subvarieties. *Ann. of Math.* **137** (1993), 407–458.

[36] Ogus, A., Hodge cycles and crystalline cohomology. In *Hodge cycles, motives, and Shimura varieties*, Lecture Notes in Math. 900, Springer-Verlag, Berlin 1982, 357–414.

[37] Poincaré, H., Sur les fonctions abéliennes. *Acta Math.* **26** (1902), 43–98.

[38] Randriambololona, H., Métriques de sous-quotient et théorème de Hilbert-Samuel arithmétique pour les faisceaux cohérents. *J. Reine Angew. Math.* **590** (2006), 67–88.

[39] Rumely, R., *Capacity theory on algebraic curves*. Lecture Notes in Math. 1378, Springer-Verlag, Berlin 1989.

[40] Siegel, C. L., Meromorphe Funktionen auf kompakten analytischen Mannigfaltigkeiten. *Nachr. Akad. Wiss. Göttingen. Math.-Phys. Kl. IIa.* **1955** (1955), 71–77.

[41] Thuillier, A., Théorie du potentiel sur les courbes en géométrie analytique non archimédienne. Applications à la théorie d'Arakelov. Thèse de Doctorat, Université de Rennes I, 2005.

[42] Viada, E., Minimal isogenies. Preprint, Technische Universität Darmstadt, 2005.

[43] Viada, E., Slopes and abelian subvariety theorem. *J. Number Theory* **112** (1) (2005), 67–115.

Université Paris-Sud, Département de Mathématiques, Bâtiment 425, 91405 Orsay Cedex, France

E-mail: jean-benoit.bost@math.u-psud.fr

Derived categories of coherent sheaves

Tom Bridgeland

Abstract. We discuss derived categories of coherent sheaves on algebraic varieties. We focus on the case of non-singular Calabi–Yau varieties and consider two unsolved problems: proving that birational varieties have equivalent derived categories, and computing the group of derived autoequivalences. We also introduce the space of stability conditions on a triangulated category and explain its relevance to these two problems.

Mathematics Subject Classification (2000). 18E30, 14J32.

Keywords. Derived categories, coherent sheaves.

1. Introduction

In the usual approach to the study of algebraic varieties one focuses directly on geometric properties of the varieties in question. Thus one considers embedded curves, hyperplane sections, branched covers and so on. A more algebraic approach is to study the varieties indirectly via their (derived) categories of coherent sheaves. This second approach has been taken up by an increasing number of researchers in the last few years. We can perhaps identify three reasons for this new emphasis on categorical methods.

Firstly, algebraic geometers have been attempting to understand string theory. The conformal field theory associated to a variety in string theory contains a huge amount of non-trivial information. However this information seems to be packaged in a categorical way rather than in directly geometric terms. This perhaps first became clear in Kontsevich's famous homological mirror symmetry conjecture [31]. Building on Kontsevich's work, it is now understood that the derived category of coherent sheaves appears in string theory as the category of branes in a topological twist of the sigma model. This means that many dualities in string theory can be described mathematically as equivalences of derived categories.

A second motivation for studying varieties via their sheaves is that this approach is expected to generalise more easily to non-commutative varieties. Although the definition of such objects is not yet clear, there are many interesting examples. In general non-commutative objects have no points in the usual sense, so that direct geometrical methods do not apply. Nonetheless one can hope that the (derived) category of coherent sheaves is well-defined and has similar properties to the corresponding object in the commutative case.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

A third reason is that recent results leave the impression that categorical methods enable one to obtain a truer description of certain varieties than current geometric techniques allow. For example many equivalences relating the derived categories of pairs of varieties are now known to exist. Any such equivalence points to a close relationship between the two varieties in question, and these relationships are often impossible to describe by other methods. Similarly, some varieties have been found to have interesting groups of derived autoequivalences, implying the existence of symmetries associated to the variety that are not visible in the geometry. It will be interesting to see whether a more geometric framework for these phenomena emerges over the next few years.

Despite a lot of recent work our understanding of derived categories of coherent sheaves is still quite primitive. To illustrate this we shall focus in this paper on two easily stated problems, both of which have been around since Bondal and Orlov's pioneering work in the nineties [5], and both of which remain largely unsolved. Of course the choice of these problems is very much a matter of personal taste. Other survey articles are available by Bondal and Orlov [7], by Hille and van den Bergh [19] and by Rouquier [45].

Of the many important contributions not discussed in detail here, we must at least mention Kuznetsov's beautiful work on semi-orthogonal decompositions for Fano varieties. As well as specific results on Fano threefolds, Kuznetsov has obtained a general theory of homological projective duality with many applications to derived categories of varieties of interest in classical algebraic geometry. The theory also leads to some extremely interesting derived equivalences between non-commutative varieties. For more details we refer the reader to the original papers [32], [33], [34], [35].

Acknowledgements. I am very grateful to Antony Maciocia for teaching me about derived categories in the first place, and to Alastair King, Miles Reid and Richard Thomas for explaining many things since then. The material in Section 5 has benefited from discussions with Yukinobu Toda. The author is supported by a Royal Society University Research Fellowship.

Notation. We write $\mathrm{D}(X) := \mathrm{D}^b \operatorname{Coh}(X)$ for the bounded derived category of coherent sheaves on a variety X. All varieties are assumed to be over the complex numbers.

2. Some basic problems

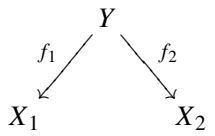
In this section we describe the basic problems we shall focus on and review some of the known results relating to them.

2.1. The following conjecture was first made by Bondal and Orlov [5].

Conjecture 2.1 (Bondal, Orlov). If X_1 and X_2 are birational smooth projective Calabi–Yau varieties of dimension n then there is an equivalence of categories $\mathrm{D}(X_1) \cong \mathrm{D}(X_2)$.

So far this is only known to hold in dimension $n = 3$. We shall describe the proof in this case in Section 3 below. There are also some local results available in all dimensions.

Theorem 2.2 (Bondal, Orlov [5]). *Suppose that X_1 and X_2 are related by a standard flop, so that there is a diagram*

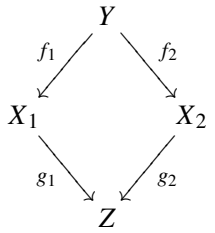


with f_i being the blow-up of a smooth subvariety $E_i \cong \mathbb{P}^d \subset X_i$ with normal bundle $\mathcal{O}(-1)^{d+1}$. Then the functor

$$\mathbf{R}f_{2,*} \circ \mathbf{L}f_1^* \colon \mathrm{D}(X_1) \to \mathrm{D}(X_2)$$

is an equivalence.

Theorem 2.3 (Kawamata [26], Namikawa [39]). *Suppose that X_1 and X_2 are related by a Mukai flop, so that there is a diagram*



with f_i being the blow-up of a smooth subvariety $E_i \cong \mathbb{P}^d \subset X_i$ with normal bundle equal to the cotangent bundle $\Omega^1(\mathbb{P}^d)$, and g_i the contraction of E_i to a point. Let $p_i \colon X_1 \times_Z X_2 \to X_i$ be the projection of the fibre product onto its factors. Then the functor

$$\mathbf{R}p_{2,*} \circ \mathbf{L}p_1^* \colon \mathrm{D}(X_1) \to \mathrm{D}(X_2)$$

is an equivalence, whilst in general the functor $\mathbf{R}f_{2,*} \circ \mathbf{L}f_1^*$ is not.

A result of Wierzba and Wisniewski [51] states that birationally equivalent complex symplectic fourfolds are related by a finite chain of Mukai flops, so Theorem 2.3 gives

Corollary 2.4. *Conjecture* 2.1 *holds when X_1 and X_2 are complex symplectic fourfolds.*

Kawamata and Namikawa have also studied certain stratified Mukai flops [29], [40]. In these cases neither a common resolution nor the fibre product give rise to an equivalence, although equivalences do nonetheless exist. Thus we see that the first problem faced in attempting to prove Conjecture 2.1 is writing down a candidate equivalence.

2.2. A famous variant of Conjecture 2.1 is the derived McKay correspondence, first stated by Reid [43], [44].

Conjecture 2.5. Let $G \subset \mathrm{SL}(n, \mathbb{C})$ be a finite subgroup. Let $\mathrm{D}_G(\mathbb{C}^n)$ denote the bounded derived category of the category of G-equivariant coherent sheaves on $\mathbb{C}^n$. Suppose $Y \to \mathbb{C}^n/G$ is a crepant resolution of singularities. Then there is an equivalence of categories $\mathrm{D}(Y) \cong \mathrm{D}_G(\mathbb{C}^n)$.

It seems certain that any method which proves Conjecture 2.1 will also apply to this case. On the other hand the McKay correspondence lends itself to more algebraic methods of attack.

Conjecture 2.5 is known to hold in dimension $n = 3$ due to work of Bridgeland, King and Reid [10] together with the results of [11]. In this case the quotient variety $\mathbb{C}^3/G$ always has a crepant resolution, a distinguished choice being given by Nakamura's G-Hilbert scheme. This scheme is perhaps best thought of as a moduli space of representations of the skew group algebra $\mathbb{C}[x, y, z] * G$ that are stable with respect to a certain choice of stability condition. Craw and Ishii [16] showed how to obtain other resolutions by varying the stability condition.

The conjecture has recently been shown to hold in two other situations, although the methods of proof are very different from that employed in the dimension three case. Firstly, Bezrukavnikov and Kaledin [4] used characteristic p methods and deformation quantization to deal with symplectic group quotients.

Theorem 2.6 (Bezrukavnikov, Kaledin [4]). *Conjecture* 2.5 *holds when G preserves a complex symplectic form on $\mathbb{C}^n$.*

Secondly, a special case of Kawamata's work [27] on the effect of toroidal flips and flops on derived categories is the following result.

Theorem 2.7 (Kawamata [28]). *Conjecture* 2.5 *holds when G is abelian.*

Note that in this case all resolutions are toric. Kawamata's proof uses the toric minimal model programme together with explicit computations of Hom spaces.

2.3. The second problem we wish to focus on is even easier to state, namely

Problem 2.8. *Compute the group of autoequivalences of the derived category of a smooth projective Calabi–Yau variety X.*

If X is any smooth projective variety there are certain standard autoequivalences of $\mathrm{D}(X)$, namely the automorphisms of X itself, twists by elements of $\mathrm{Pic}(X)$, and the shift functor [1]. Bondal and Orlov [5], [6] showed that if X has ample canonical or anticanonical bundle, then these standard equivalences generate $\mathrm{Aut}\,\mathrm{D}(X)$. Thus the group $\mathrm{Aut}\,\mathrm{D}(X)$ is not so interesting. In fact $\mathrm{Aut}\,\mathrm{D}(X)$ seems to be most interesting when X is Calabi–Yau, so we shall concentrate on this case.

Certain types of objects are known to define autoequivalences. The most well-known are spherical objects.

Theorem 2.9 (Seidel, Thomas [46]). *Let X be a smooth projective Calabi–Yau of dimension n and let $S \in \mathrm{D}(X)$ be a spherical object, which is to say an object satisfying*

$$\mathrm{Hom}^k_{\mathrm{D}(X)}(S, S) = \begin{cases} \mathbb{C} & \text{if } k = 0 \text{ or } n, \\ 0 & \text{otherwise.} \end{cases}$$

Then there is an autoequivalence $\Phi_S \in \mathrm{Aut}(\mathrm{D}(X))$ such that for any object $E \in \mathrm{D}(X)$ there is an exact triangle

$$\mathrm{Hom}_{\mathrm{D}(X)}(S, E) \otimes S \longrightarrow E \longrightarrow \Phi_S(E).$$

Horja [20] generalised the construction of Seidel and Thomas to define new autoequivalences. Szendrői [47] showed how certain configurations of surfaces in Calabi–Yau threefolds lead to interesting groups of autoequivalences. More recently, Huybrechts and Thomas [21] have studied projective space objects, which also give rise to autoequivalences.

Problem 2.8 is already non-trivial when X is an elliptic curve. Since X can be identified with its dual abelian variety $\mathrm{Pic}^0(X)$ in the dimension one case, the original Fourier transform of Mukai [37] defines an autoequivalence $\mathcal{F} \in \mathrm{Aut}\,\mathrm{D}(X)$. This enables one to prove

Theorem 2.10. *The group $\mathrm{Aut}\,\mathrm{D}(X)$ is generated by the standard autoequivalences together with $\mathcal{F}$. There is an exact sequence*

$$1 \longrightarrow \mathbb{Z} \times (\mathrm{Aut}(X) \ltimes \mathrm{Pic}^0(X)) \longrightarrow \mathrm{Aut}\,\mathrm{D}(X) \xrightarrow{f} \mathrm{SL}(2, \mathbb{Z}) \longrightarrow 1.$$

where the factor of $\mathbb{Z}$ is generated by the double shift [2]*, and the map f is defined by*

$$\begin{pmatrix} \mathrm{rank}(\Phi(E)) \\ \deg(\Phi(E)) \end{pmatrix} = f(\Phi) \begin{pmatrix} \mathrm{rank}(E) \\ \deg(E) \end{pmatrix}$$

for all $E \in \mathrm{D}(X)$.

If L is a degree one line bundle on X then, neglecting shifts, the autoequivalences $- \otimes L$ and $\mathcal{F}$ generate a subgroup of $\mathrm{Aut}\,\mathrm{D}(X)$ isomorphic to $\mathrm{SL}(2, \mathbb{Z})$. This $\mathrm{SL}(2, \mathbb{Z})$ action discovered by Mukai was perhaps an early pointer to homological mirror symmetry.

Theorem 2.10 was generalised by Orlov [42] who calculated the group $\mathrm{Aut}\,\mathrm{D}(X)$ for all abelian varieties X. The group of derived autoequivalences of the minimal resolution of a Kleinian singularity was studied in [23]. But beyond these results not much is known. We shall consider the case when X is a K3 surface in Section 6.

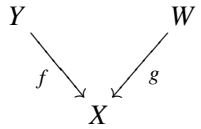
3. Threefold flops

In this section we shall be concerned with the following result.

Theorem 3.1. *Conjecture* 2.1 *holds when X_1 and X_2 have dimension three.*

Of course in dimensions two or less, birational Calabi–Yau varieties are isomorphic. We start by giving a bare outline of the original proof of Theorem 3.1 and then return to make some further remarks on each step. Full details of the proof can be found in [11]. A more direct proof has since been found by Kawamata [26] using results of Van den Bergh [50]. Nonetheless, the original proof is worth describing since it gives a moduli-theoretic interpretation of threefold flops which is interesting in its own right.

Sketch of proof. The first step is to apply the minimal model programme in the shape of a result of Kawamata [24] (a simpler proof is given in Kollár [30]) which says that any birational transform between smooth projective threefolds that preserves the canonical class can be split into a finite sequence of flops. Thus we can reduce to the case where we have a flopping diagram



Here Y and W are smooth and projective, X has Gorenstein terminal singularities, and f and g are crepant birational maps contracting only finitely many rational curves.

The second step is to define a full subcategory $\mathrm{Per}(Y/X) \subset \mathrm{D}(Y)$ consisting of objects $E \in \mathrm{D}(Y)$ satisfying the following three conditions

(a) $H^i(E) = 0$ unless $i = 0$ or -1,

(b) $\mathbf{R}^1 f_*\, H^0(E) = 0$ and $\mathbf{R}^0 f_*\, H^{-1}(E) = 0$,

(c) $\mathrm{Hom}_Y(H^0(E), C) = 0$ for any sheaf C on Y satisfying $\mathbf{R} f_*(C) = 0$.

In fact $\mathrm{Per}(Y/X)$ is the heart of a bounded t-structure on $\mathrm{D}(Y)$ and hence is abelian. We call the objects of $\mathrm{Per}(Y/X)$ perverse coherent sheaves.

Define a perverse point sheaf on Y to be an object of $\mathrm{Per}(Y/X)$ which has the same Chern classes as the structure sheaf of a point $y \in Y$, and which is a quotient

of $\mathcal{O}_Y$ in $\mathrm{Per}(Y/X)$. The third step is to show that there is a fine moduli space M for perverse point sheaves on Y. Since perverse point sheaves are not sheaves in general, but complexes of sheaves with nontrivial cohomology in more than one place, the usual moduli space techniques do not apply. In fact we can side-step this issue as follows.

By definition each perverse point sheaf E fits into a short exact sequence

$$0 \longrightarrow F \longrightarrow \mathcal{O}_Y \longrightarrow E \longrightarrow 0$$

in $\mathrm{Per}(Y/X)$. Taking the long exact sequence in cohomology shows that F is actually a sheaf, although not necessarily torsion-free. One can construct a fine moduli space M for the objects F using a standard GIT argument, and since the Es and Fs determine one another this space is also a fine moduli space for perverse point sheaves.

The moduli space M has a natural map to X since it is easily shown that if E is a perverse point sheaf on Y then $\mathbf{R}f_*(E)$ is the structure sheaf of a point of X. The fourth step is to show that M is smooth and that the universal family $\mathcal{P}$ on $M \times Y$ induces a Fourier–Mukai equivalence $\Phi\colon \mathrm{D}(M) \to \mathrm{D}(Y)$. This involves an application of the famous new intersection theorem. Once one has this it is easy to see that M with its natural map to X is the flop of $Y \to X$ and hence can be identified with W.

Finally, we should note that Chen [15, Prop. 4.2] was able to show that the universal family $\mathcal{P}$ is isomorphic to the object $\mathcal{O}_{W\times_X Y}$. Thus the resulting functor Φ is of the same form used in Bondal and Orlov's result Theorem 2.2. Later Kawamata [26] was able to show directly that this functor gives an equivalence. □

Example 3.2. To understand why the moduli of perverse point sheaves gives the flop consider the simplest example when $f\colon Y \to X$ is the contraction of a non-singular rational curve C with normal bundle $\mathcal{O}_C(-1) \oplus \mathcal{O}_C(-1)$.

Structure sheaves of points $y \in Y$ are objects of the category $\mathrm{Per}(Y/X)$ for all $y \in Y$. However, if $y \in C$ then the nonzero map $\mathcal{O}_C(-1) \to \mathcal{O}_y$ means that $\mathcal{O}_y$ is not an object of $\mathrm{Per}(Y/X)$, so that $\mathcal{O}_y$ is not a quotient of $\mathcal{O}_Y$ in $\mathrm{Per}(Y/X)$ and hence is not a perverse point sheaf. In fact for $y \in C$, the sheaf $\mathcal{O}_y$ fits into the exact sequence

$$0 \longrightarrow \mathcal{O}_C(-1) \longrightarrow \mathcal{O}_C \longrightarrow \mathcal{O}_y \longrightarrow 0. \tag{1}$$

Now $\mathcal{O}_C$ is a perverse sheaf, but $\mathcal{O}_C(-1)$ is not, so that the triangle in $\mathrm{D}(Y)$ arising from (1) does not define an exact sequence in $\mathrm{Per}(Y/X)$. However the complex obtained by shifting $\mathcal{O}_C(-1)$ to the left by one place *is* a perverse sheaf, so there is an exact sequence of perverse sheaves

$$0 \longrightarrow \mathcal{O}_C \longrightarrow \mathcal{O}_y \longrightarrow \mathcal{O}_C(-1)[1] \longrightarrow 0 \tag{2}$$

which could be thought of as destabilizing $\mathcal{O}_y$.

Flipping the extension of perverse sheaves (2) gives objects of $\mathrm{Per}(Y/X)$ fitting into an exact sequence of perverse sheaves

$$0 \longrightarrow \mathcal{O}_C(-1)[1] \longrightarrow E \longrightarrow \mathcal{O}_C \longrightarrow 0. \tag{3}$$

It is easy to see that these objects E are perverse point sheaves. Note that they are not sheaves, indeed any such object has two nonzero cohomology sheaves $H^{-1}(E) = \mathcal{O}_C(-1)$ and $H^0(E) = \mathcal{O}_C$. Roughly speaking, the space W is obtained from X by replacing the rational curve C parameterising extensions (2) by another rational curve C' parameterising extensions (3).

We now make some further remarks on the proof with an eye to generalising the method to higher dimensions, although, as we shall see, the prospects for doing this do not seem particularly good.

It is clear that the key step in the proof is the introduction of the category $\mathrm{Per}(Y/X)$. There are two possible ways to arrive at this definition as we shall now explain. Unfortunately neither of them seems to generalise directly to higher dimensional situations.

Construction 1. The first approach is to use the theory of tilting in abelian categories as developed by Happel, Reiten and Smalø [18]. This is an abstraction of the notion of tilting of algebras due to Brenner and Butler [9]. In fact in our situation the full subcategory

$$\mathcal{F} = \{E \in \mathrm{Coh}(Y) : f_*(E) = 0\}$$

is the torsion-free subcategory of a torsion pair on $\mathrm{Coh}(Y)$. The corresponding torsion subcategory consists of objects $E \in \mathrm{Coh}(Y)$ for which $\mathbf{R}^1 f_*(E) = 0$ and $\mathrm{Hom}_Y(E, C) = 0$ for any sheaf C on Y satisfying $\mathbf{R} f_*(C) = 0$. Tilting with respect to this torsion pair gives the t-structure on $\mathrm{D}(Y)$ whose heart is $\mathrm{Per}(Y/X)$.

Construction 2. Van den Bergh [50] found another characterisation of $\mathrm{Per}(Y/X)$ involving sheaves of non-commutative algebras. He first introduced a particular locally-free sheaf P on Y and considered the coherent sheaf of algebras $\mathcal{A} = \mathbf{R} f_* \mathcal{E}\mathrm{nd}_{\mathcal{O}_Y}(P)$ on X. Then he showed that there is an equivalence

$$\mathbf{R} f_* \mathcal{H}\mathrm{om}_{\mathcal{O}_Y}(P, -) \colon \mathrm{D}(Y) \longrightarrow \mathrm{D}(\mathcal{A}),$$

where $\mathrm{D}(\mathcal{A})$ is the bounded derived category of the abelian category $\mathrm{Coh}(\mathcal{A})$ of coherent sheaves of $\mathcal{A}$-modules on X. Pulling back the standard t-structure on $\mathrm{D}(\mathcal{A})$ gives the t-structure on $\mathrm{D}(Y)$ whose heart is $\mathrm{Per}(Y/X)$. In particular the above derived equivalence restricts to give an equivalence of abelian subcategories $\mathrm{Per}(Y/X) \to \mathrm{Coh}(\mathcal{A})$.

The third step in the proof of Theorem 3.1 was to construct a fine moduli space for perverse point sheaves. The trick we used in dimension three will not work in higher dimensions. Instead one must construct the moduli space directly. Recent results of Inaba [22] and Toen and Vaquie [49] construct spaces parameterising very general classes of objects of derived categories. For applications we need a more geometric approach. We first need to understand what it means for an object of a triangulated category to be stable. The general notion of a stability condition introduced in the next section sheds some light on this. After that we need to construct good moduli spaces of stable objects, where good might mean for example that the moduli space

is a projective scheme. We refer to Abramovich and Polishchuk [1] for some recent progress in this direction.

The final step in the proof of Theorem 3.1 was very much dependent on the dimension three assumption. The estimates needed to apply the intersection theorem just do not work in higher dimensions. In fact, in dimension four and higher, one can obtain singular varieties by flopping non-singular ones, so we cannot expect to prove that the relevant moduli space is smooth. Derived categories of singular varieties are still poorly understood at present, a point to which we shall return in the last section.

4. Stability conditions

The notion of a stability condition was introduced in [12] as a way to understand Douglas' work on π-stability for D-branes in string theory [17]. Here we wish to emphasise the purely mathematical aspects of this definition.

In the context of the present article stability conditions are relevant for three reasons. Firstly, the choice of a stability condition picks out classes of stable objects for which one can hope to form well-behaved moduli spaces. Secondly the space of all stability conditions Stab(D) allows one to bring geometric methods to bear on the problem of understanding t-structures on D. Finally, the space Stab(D) provides a complex manifold on which the group Aut(D) naturally acts.

In this section D is a fixed triangulated category. We shall assume that D is essentially small, i.e. equivalent to a category whose objects form a set. The Grothendieck group of D is denoted $K(\mathrm{D})$. For details on the results of this section see [12].

4.1. The definition of a stability condition is as follows.

Definition 4.1. A stability condition $\sigma = (Z, \mathcal{P})$ on D consists of a group homomorphism $Z\colon K(\mathrm{D}) \to \mathbb{C}$ called the central charge, and full additive subcategories $\mathcal{P}(\phi) \subset \mathrm{D}$ for each $\phi \in \mathbf{R}$, satisfying the following axioms:

(a) if $E \in \mathcal{P}(\phi)$ then $Z(E) = m(E)\exp(i\pi\phi)$ for some $m(E) \in \mathbf{R}_{>0}$,

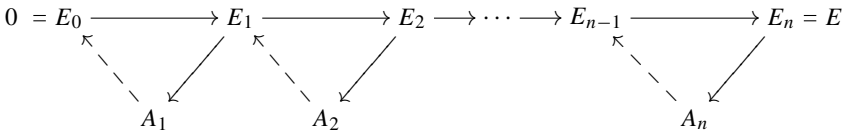
(b) for all $\phi \in \mathbf{R}$, $\mathcal{P}(\phi + 1) = \mathcal{P}(\phi)[1]$,

(c) if $\phi_1 > \phi_2$ and $A_j \in \mathcal{P}(\phi_j)$ then $\mathrm{Hom}_{\mathrm{D}}(A_1, A_2) = 0$,

(d) for each nonzero object $E \in \mathrm{D}$ there is a finite sequence of real numbers

$$\phi_1 > \phi_2 > \cdots > \phi_n$$

and a collection of triangles



with $A_j \in \mathcal{P}(\phi_j)$ for all j.

Given a stability condition $\sigma = (Z, \mathcal{P})$ as in the definition, each subcategory $\mathcal{P}(\phi)$ is abelian. The nonzero objects of $\mathcal{P}(\phi)$ are said to be semistable of phase ϕ in σ, and the simple objects of $\mathcal{P}(\phi)$ are said to be stable. It follows from the other axioms that the decomposition of an object $0 \neq E \in \mathrm{D}$ given by axiom (d) is uniquely defined up to isomorphism. Write $\phi^+_\sigma(E) = \phi_1$ and $\phi^-_\sigma(E) = \phi_n$. The mass of E is defined to be the positive real number $m_\sigma(E) = \sum_i |Z(A_i)|$.

For any interval $I \subset \mathbb{R}$, define $\mathcal{P}(I)$ to be the extension-closed subcategory of D generated by the subcategories $\mathcal{P}(\phi)$ for $\phi \in I$. Thus, for example, the full subcategory $\mathcal{P}((a, b))$ consists of the zero objects of D together with those objects $0 \neq E \in \mathrm{D}$ which satisfy $a < \phi^-_\sigma(E) \leqslant \phi^+_\sigma(E) < b$. A stability condition is called locally finite if there is some $\varepsilon > 0$ such that each quasi-abelian category $\mathcal{P}((\phi - \varepsilon, \phi + \varepsilon))$ is of finite length.

4.2. If $\sigma = (Z, \mathcal{P})$ is a stability condition on D, one can check that the full subcategory $\mathcal{A} = \mathcal{P}((0, 1])$ is the heart of a bounded t-structure on D. We call it the heart of σ. The stability condition σ can easily be reconstructed from its heart $\mathcal{A}$ together with the central charge map Z. Since this will be important in Section 5 we spell it out in a little more detail.

Define a stability function on an abelian category $\mathcal{A}$ to be a group homomorphism $Z\colon K(\mathcal{A}) \to \mathbb{C}$ such that

$$0 \neq E \in \mathcal{A} \implies Z(E) \in \mathbf{R}_{>0} \exp(i\pi\phi(E)) \text{ with } 0 < \phi(E) \leqslant 1.$$

The real number $\phi(E) \in (0, 1]$ is called the phase of the object E.

A nonzero object $E \in \mathcal{A}$ is said to be semistable with respect to Z if every subobject $0 \neq A \subset E$ satisfies $\phi(A) \leqslant \phi(E)$. The stability function Z is said to have the Harder–Narasimhan property if every nonzero object $E \in \mathcal{A}$ has a finite filtration

$$0 = E_0 \subset E_1 \subset \cdots \subset E_{n-1} \subset E_n = E$$

whose factors $F_j = E_j/E_{j-1}$ are semistable objects of $\mathcal{A}$ with

$$\phi(F_1) > \phi(F_2) > \cdots > \phi(F_n).$$

Given a stability condition on D, the central charge defines a stability function on its heart $\mathcal{A} = \mathcal{P}((0, 1]) \subset \mathrm{D}$, and the decompositions of axiom (d) give Harder–Narasimhan filtrations. Conversely, given a bounded t-structure on D together with a stability function Z on its heart $\mathcal{A} \subset \mathrm{D}$, we can define subcategories $\mathcal{P}(\phi) \subset \mathcal{A} \subset \mathrm{D}$ to be the semistable objects in $\mathcal{A}$ of phase ϕ for each $\phi \in (0, 1]$. Axiom (b) then fixes $\mathcal{P}(\phi)$ for all $\phi \in \mathbf{R}$. Thus we have

Proposition 4.2. *To give a stability condition on* D *is equivalent to giving a bounded t-structure on* D *and a stability function on its heart with the Harder–Narasimhan property.*

4.3. The set Stab(D) of locally-finite stability conditions on D has a natural topology induced by the metric

$$d(\sigma_1,\sigma_2)=\sup_{0\neq E\in \mathrm{D}}\left\{|\phi^-_{\sigma_2}(E)-\phi^-_{\sigma_1}(E)|,\,|\phi^+_{\sigma_2}(E)-\phi^+_{\sigma_1}(E)|,\,\left|\log\frac{m_{\sigma_2}(E)}{m_{\sigma_1}(E)}\right|\right\}\in[0,\infty].$$

The following result was proved in [11]. Its slogan is that deformations of the central charge lift uniquely to deformations of the stability condition.

Theorem 4.3. *For each connected component* $\Sigma\subset \mathrm{Stab}(\mathrm{D})$ *there is a linear subspace* $V(\Sigma)\subset \mathrm{Hom}_{\mathbb{Z}}(K(\mathrm{D}),\mathbb{Z})$ *with a well-defined linear topology and a local homeomorphism* $\mathcal{Z}\colon \Sigma\to V(\Sigma)$ *which maps a stability condition* $(Z,\mathcal{P})$ *to its central charge* Z.

If X is a smooth projective variety we write $\mathrm{Stab}(X)$ for the set of locally-finite stability conditions on $\mathrm{D}(X)$ for which the central charge Z factors via the Chern character map $\mathrm{ch}\colon K(X)\to H^*(X,\mathbb{Q})$. Theorem 4.3 immediately implies that $\mathrm{Stab}(X)$ is a finite-dimensional complex manifold. By Proposition 4.2, the points of $\mathrm{Stab}(X)$ parameterise bounded t-structures on $\mathrm{D}(X)$, together with the extra data of the map Z.

So far the manifolds $\mathrm{Stab}(X)$ have only been computed for varieties X of dimension one [36], [41]. The case of K3 and abelian surfaces was studied in [13]; we shall return to the K3 case in Section 6. Various non-compact examples have also been considered (see e.g. [14], [48]).

5. Stability conditions and threefold flops

Let $f\colon Y\to X$ be a small, crepant birational map of threefolds with Y smooth. Thus X has terminal Gorenstein singularities and f contracts a finite number of rational curves C, each satisfying $K_Y\cdot C=0$. The aim of this section is to show how the category of perverse sheaves defined in Section 3 arises naturally by considering stability conditions on Y. We shall go into more detail than we have in previous sections, since the results provide an excellent example of how stability conditions can be used to relate the derived category to geometry. For full proofs the reader can consult Toda's paper [48]. The string theory point of view on the same material is described in [2].

5.1. Define $\mathrm{D}(Y/X)$ to be the full subcategory of $\mathrm{D}(Y)$ consisting of objects supported on (a fat neighbourhood of) the exceptional locus of f. Since $\mathrm{D}(Y/X)$ only depends on a formal neighbourhood of the singular locus of X we may as well assume that X is the spectrum of a complete local ring.

One can easily show that there is an isomorphism of abelian groups

$$K(\mathrm{D}(Y/X))\cong N_1(Y/X)\oplus\mathbb{Z}.$$

It will be convenient to consider the codimension one slice of the full space $\mathrm{Stab}(\mathrm{D}(Y/X))$ consisting of stability conditions satisfying the additional condition that $Z(\mathcal{O}_y) = -1$, where $\mathcal{O}_y$ is the structure sheaf of a point $y \in Y$. We shall denote it simply by $\mathrm{Stab}(Y/X)$. Since the dual of $N_1(Y/X)$ is naturally $N^1(Y/X)$, the central charge $Z\colon K(\mathrm{D}(Y/X)) \to \mathbb{C}$ of a stability condition in $\mathrm{Stab}(Y/X)$ is defined by an element $\beta + i\omega$ of the vector space $N^1(Y/X) \otimes \mathbb{C}$. Explicitly the correspondence is given by

$$Z(E) = (\beta + i\omega) \cdot \mathrm{ch}_2(E) - \mathrm{ch}_3(E). \qquad (\dagger)$$

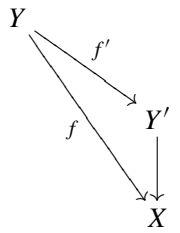
The map $\mathcal{Z}$ of Theorem 4.3 now becomes a map

$$\mathcal{Z}\colon \mathrm{Stab}(Y/X) \longrightarrow N^1(X/Y) \otimes \mathbb{C}.$$

The standard t-structure on $\mathrm{D}(Y)$ induces a bounded t-structure on $\mathrm{D}(Y/X)$ whose heart $\mathrm{Coh}(Y/X) = \mathrm{Coh}(Y) \cap \mathrm{D}(Y/X)$ is the subcategory of $\mathrm{Coh}(Y)$ consisting of sheaves supported on the exceptional locus. A simple application of Proposition 4.2 gives

Proposition 5.1. *A stability function for the t-structure with heart* $\mathrm{Coh}(Y/X) \subset \mathrm{D}(Y/X)$ *is given by* (†) *with* $\beta, \omega \in N^1(Y/X) \otimes \mathbf{R}$ *such that* $\omega \in \mathrm{Amp}(Y/X)$ *lies in the ample cone. This stability function has the Harder–Narasimhan property. Thus there is a region in* $U(Y/X) \subset \mathrm{Stab}(Y/X)$ *isomorphic to the complexified ample cone of* Y/X.

5.2. Suppose now that $\sigma = (Z, \mathcal{P})$ is a stability condition in the boundary of the region $U(Y/X)$. Then Z is given by the formula above for some $\beta, \omega \in N^1(Y/X) \otimes \mathbf{R}$ with ω an f-nef divisor which is not ample. There is a unique birational map $f'\colon Y \to Y'$ factoring $f\colon Y \to X$ such that $\omega = (f')^*(\omega')$ with ω' an ample $\mathbf{R}$-divisor on Y'.



The map f' is also crepant and contracts precisely those irreducible rational curves $C \subset Y$ satisfying $\omega \cdot C = 0$.

Let C be a rational curve contracted by f'. Note that for all stability conditions in $U(Y/X)$, the objects $\mathcal{O}_C(k)$ are stable, since their only proper quotients in $\mathrm{Coh}(Y/X)$ are skyscraper sheaves. It follows that these objects are at least semistable in σ, and hence $Z(\mathcal{O}_C(k)) \neq 0$. Since $\omega \cdot C = 0$ we must have $\beta \cdot C \notin \mathbb{Z}$.

The map Z defined by (†) no longer defines a stability function for $\mathrm{Coh}(Y/X)$ because if C is contracted by f' and $k \ll 0$ then $Z(\mathcal{O}_C(k))$ lies on the positive real axis. Thus the heart of the stability condition σ cannot be $\mathrm{Coh}(Y/X) \subset \mathrm{D}(Y/X)$.

Instead, consider the perverse t-structure on $\mathrm{D}(Y)$ induced by the contraction f' and restrict it to $\mathrm{D}(Y/X)$ to give a bounded t-structure whose heart $\mathrm{Per}(Y/Y') \cap \mathrm{D}(Y/X)$ we shall also denote $\mathrm{Per}(Y/Y')$. One can then show that providing $0 < \beta \cdot C < 1$ for all irreducible curves C with $\omega \cdot C = 0$ then the heart of σ is indeed equal to $\mathrm{Per}(Y/Y')$.

Suppose for a moment that σ lies on a codimension one face of the boundary of $U(Y/X)$, by which we mean that $\mathrm{Pic}(Y/Y') = \mathbb{Z}$. Note that $\mathrm{Pic}(Y/X)$ acts on $\mathrm{Stab}(Y/X)$ and on the region $U(Y/X)$. Thus twisting by a line bundle we can always assume that the condition $0 < \beta \cdot C < 1$ holds.

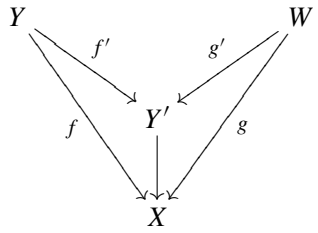
Conversely, considering the definition of $\mathrm{Per}(Y/Y')$ as a tilt, one can show that if the above condition on β and ω holds then the map Z defined by (†) gives a stability function on $\mathrm{Per}(Y/Y')$, and this is easily checked to satisfy the Harder–Narasimhan property. These arguments give

Proposition 5.2. *Let $f' \colon Y \to Y'$ be a birational map factoring $f \colon Y \to X$. A stability function for the t-structure with heart* $\mathrm{Per}(Y/Y')$ *is given by* (†) *for classes $\beta, \omega \in N^1(Y/X) \otimes \mathbf{R}$ such that $\omega = (f')^*(\omega')$ for some $\omega' \in \mathrm{Amp}(Y'/X)$ and $0 < \beta \cdot C < 1$ for all curves C contracted by f'. The resulting stability conditions lie in the boundary of the region $U(Y/X)$. Conversely, any stability condition lying on a codimension one face of the boundary of $U(Y/X)$ is of this form up to a twist by a line bundle.*

The key point we wish to emphasize is that the perverse t-structure arises naturally by passing to the boundary of the region $U(Y/X)$.

5.3. Consider now an open neighbourhood of a point σ lying in the boundary of $U(Y/X)$. This region will contain stability conditions with central charges Z defined by (†) with ω lying in the ample cone of a different birational model of $f \colon Y \to X$. It can be shown that all these stability conditions can be obtained by pulling back stability conditions of the form given in Proposition 5.1 by derived equivalences.

For example, suppose $g \colon W \to X$ is obtained from f by taking the flop of a contraction $f' \colon Y \to Y'$ factoring f with $\mathrm{Pic}(Y/Y') \cong \mathbb{Z}^{\oplus d}$.



Such a map f' defines a codimension d face of the ample cone $\mathrm{Amp}(Y/X)$. Let $\Phi \colon \mathrm{D}(Y) \longrightarrow \mathrm{D}(W)$ be the equivalence with kernel $\mathcal{O}_{Y \times_{Y'} W}$ on $Y \times W$. Then Φ

induces an isomorphism fitting into a diagram

$$\begin{array}{ccc} \mathrm{Stab}(Y/X) & \xrightarrow{\quad\Phi\quad} & \mathrm{Stab}(W/X) \\ {\scriptstyle\mathcal{Z}}\downarrow & & \downarrow{\scriptstyle\mathcal{Z}} \\ N^1(Y/X)\otimes\mathbb{C} & \xrightarrow{(g^{-1}\circ f)_*} & N^1(W/X)\otimes\mathbb{C} \end{array}$$

where the isomorphism $N^1(Y/X) \to N^1(W/X)$ is induced by the codimension one isomorphism $g^{-1} \circ f \colon Y \dashrightarrow W$. One can show that the closure of the inverse image $\Phi^{-1}(U(W/X))$ intersects the closure of $U(Y/X)$ along a real codimension d component of the boundary.

It is well known that the general hyperplane section of X is a Kleinian ADE surface singularity. Furthermore the abelian group $N_1(Y/X)$ can be identified with the root lattice of the corresponding simple Lie algebra. Let $\Lambda \subset N_1(Y/X)$ be the set of roots.

Theorem 5.3 (Toda, [48]). *The map*

$$\mathcal{Z}\colon\ \mathrm{Stab}(Y/X) \longrightarrow N^1(Y/X)\otimes\mathbb{C}$$

restricted to the connected component of $\mathrm{Stab}(Y/X)$ *containing* $U(Y/X)$ *is a covering map of the hyperplane complement*

$$\{\beta + i\omega \in N^1(Y/X)\otimes\mathbb{C} : \beta + i\omega \cdot C \notin \mathbb{Z} \text{ for all } C \in \Lambda\}.$$

For each smooth birational model $W \to X$ *and each Fourier–Mukai equivalence*

$$\Phi\colon\ \mathrm{D}(Y) \longrightarrow \mathrm{D}(W)$$

over X *there is a region* $\Phi^{-1}(U(W/X)) \subset \mathrm{Stab}(Y/X)$ *isomorphic to the complexified ample cone of* W/X. *The closures of these regions cover a connected component of* $\mathrm{Stab}(Y/X)$, *and any two of the regions are either disjoint or equal.*

6. Stability conditions on K3 surfaces

Suppose that X is an algebraic K3 surface over $\mathbb{C}$. In this section we consider stability conditions on $\mathrm{D}(X)$ and explain how they may help to determine the group of autoequivalences $\mathrm{Aut}\,\mathrm{D}(X)$. Full details can be found in [13].

6.1. Following Mukai [38], one introduces the extended cohomology lattice of X by using the formula

$$\big((r_1, D_1, s_1), (r_2, D_2, s_2)\big) = D_1 \cdot D_2 - r_1 s_2 - r_2 s_1$$

to define a symmetric bilinear form on the cohomology ring

$$H^*(X,\mathbb{Z}) = H^0(X,\mathbb{Z}) \oplus H^2(X,\mathbb{Z}) \oplus H^4(X,\mathbb{Z}).$$

The resulting lattice $H^*(X, \mathbb{Z})$ is even and non-degenerate and has signature $(4, 20)$. Let $H^{2,0}(X) \subset H^2(X, \mathbb{C})$ denote the one-dimensional complex subspace spanned by the class of a nonzero holomorphic two-form Ω on X. An isometry

$$\varphi\colon H^*(X, \mathbb{Z}) \to H^*(X, \mathbb{Z})$$

is called a Hodge isometry if $\varphi \otimes \mathbb{C}$ preserves this subspace. The group of Hodge isometries of $H^*(X, \mathbb{Z})$ will be denoted $\operatorname{Aut} H^*(X, \mathbb{Z})$.

The Mukai vector of an object $E \in \mathrm{D}(X)$ is the element of the sublattice

$$\mathcal{N}(X) = \mathbb{Z} \oplus \mathrm{NS}(X) \oplus \mathbb{Z} = H^*(X, \mathbb{Z}) \cap \Omega^\perp \subset H^*(X, \mathbb{C})$$

defined by the formula

$$v(E) = (r(E), c_1(E), s(E)) = \mathrm{ch}(E)\sqrt{\mathrm{Td}(X)} \in H^*(X, \mathbb{Z}),$$

where $\mathrm{ch}(E)$ is the Chern character of E and $s(E) = \mathrm{ch}_2(E) + r(E)$. The Mukai bilinear form makes $\mathcal{N}(X)$ into an even lattice of signature $(2, \rho)$, where $1 \leqslant \rho \leqslant 20$ is the Picard number of X. The Riemann–Roch theorem shows that this form is the negative of the Euler form, that is, for any pair of objects E and F of $\mathrm{D}(X)$

$$\chi(E, F) = \sum_i (-1)^i \dim_{\mathbb{C}} \operatorname{Hom}^i_X(E, F) = -(v(E), v(F)).$$

A result of Orlov [42], extending work of Mukai [38], shows that every exact autoequivalence of $\mathrm{D}(X)$ induces a Hodge isometry of the lattice $H^*(X, \mathbb{Z})$. Thus there is a group homomorphism

$$\varpi\colon \operatorname{Aut} \mathrm{D}(X) \longrightarrow \operatorname{Aut} H^*(X, \mathbb{Z}).$$

The kernel of this homomorphism will be denoted $\operatorname{Aut}^0 \mathrm{D}(X)$.

For any point $\sigma = (Z, \mathcal{P}) \in \mathrm{Stab}(X)$ the central charge $Z(E)$ depends only on the Chern character of E and hence can be written in the form

$$Z(E) = (\mho, v(E))$$

for some vector $\mho \in \mathcal{N}(X) \otimes \mathbb{C}$. Thus the map $\mathcal{Z}$ of Theorem 4.3 becomes a map

$$\mathcal{Z}\colon \mathrm{Stab}(X) \to \mathcal{N}(X) \otimes \mathbb{C}$$

sending a stability condition to the corresponding vector $\mho \in \mathcal{N}(X) \otimes \mathbb{C}$.

Define an open subset

$$\mathcal{P}^{\pm}(X) \subset \mathcal{N}(X) \otimes \mathbb{C}$$

consisting of those vectors which span positive definite two-planes in $\mathcal{N}(X) \otimes \mathbf{R}$. This space has two connected components that are exchanged by complex conjugation. Let $\mathcal{P}^+(X)$ denote the component containing vectors of the form $\exp(i\omega) = (1, i\omega, -\omega^2/2)$ for ample divisor classes $\omega \in \mathrm{NS}(X) \otimes \mathbf{R}$. Set

$$\Delta(X) = \{\delta \in \mathcal{N}(X) : (\delta, \delta) = -2\}$$

and for each $\delta \in \Delta(X)$ let

$$\delta^{\perp} = \{\mho \in \mathcal{N}(X) \otimes \mathbb{C} : (\mho, \delta) = 0\} \subset \mathcal{N}(X) \otimes \mathbb{C}$$

be the corresponding complex hyperplane. The following result was proved in [13].

Theorem 6.1. *There is a connected component $\Sigma(X) \subset \mathrm{Stab}(X)$ that is mapped by $\mathcal{Z}$ onto the open subset*

$$\mathcal{P}_0^+(X) = \mathcal{P}^+(X) \setminus \bigcup_{\delta \in \Delta(X)} \delta^{\perp} \subset \mathcal{N}(X) \otimes \mathbb{C}.$$

Moreover, the induced map $\mathcal{Z}\colon \Sigma(X) \to \mathcal{P}_0^+(X)$ is a regular covering map and the subgroup of $\mathrm{Aut}^0 \mathrm{D}(X)$ which preserves the connected component $\Sigma(X)$ acts freely on $\Sigma(X)$ and is the group of deck transformations of $\mathcal{Z}$.

Unfortunately, Theorem 6.1 is not enough to determine the group $\mathrm{Aut}\,\mathrm{D}(X)$. Nonetheless it provides a hope to solve this problem by studying the geometry of $\mathrm{Stab}(X)$. The obvious thing to conjecture is as follows.

Conjecture 6.2. The action of $\mathrm{Aut}\,\mathrm{D}(X)$ on $\mathrm{Stab}(X)$ preserves the connected component $\Sigma(X)$. Moreover $\Sigma(X)$ is simply-connected.

This conjecture would imply that there is a short exact sequence of groups

$$1 \longrightarrow \pi_1 \mathcal{P}_0^+(X) \longrightarrow \mathrm{Aut}\,\mathrm{D}(X) \xrightarrow{\varpi} \mathrm{Aut}^+ H^*(X, \mathbb{Z}) \longrightarrow 1$$

where

$$\mathrm{Aut}^+ H^*(X, \mathbb{Z}) \subset \mathrm{Aut}\, H^*(X, \mathbb{Z})$$

is the index 2 subgroup consisting of elements which do not exchange the two components of $\mathcal{P}^{\pm}(X)$.

As a final remark in this section note that Borcherds' work on modular forms [8] allows one to write down product expansions for holomorphic functions on $\mathrm{Stab}(X)$ that are invariant under the group $\mathrm{Aut}\,\mathrm{D}(X)$. It would be interesting to connect these formulae with counting invariants for stable objects in $\mathrm{D}(X)$.

7. Derived categories and the minimal model programme

In this section we briefly mention some further recent work on birational geometry and derived categories of coherent sheaves and give some references.

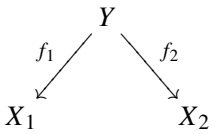
7.1. Conjecture 2.1 is the simplest version of a much more general set of conjectures about derived categories and birational geometry. The basic idea, due to Bondal and Orlov, is that each of the operations in the minimal model programme should induce fully faithful embeddings of derived categories. Thus for example, if $f\colon Y \to X$

is a birational map of smooth projective varieties, then $\mathbf{R}f_*(\mathcal{O}_Y) = \mathcal{O}_X$, and the projection formula implies that the functor

$$\mathbf{L}f^*\colon \mathrm{D}(X) \longrightarrow \mathrm{D}(Y)$$

is full and faithful. Much more generally Bondal and Orlov conjectured

Conjecture 7.1 (Bondal, Orlov). Suppose



are birational maps of smooth projective varieties. Suppose the divisor

$$K_{X_2/X_1} = f_2^*(K_{X_2}) - f_1^*(K_{X_1})$$

is effective. Then there is a fully faithful functor $F\colon \mathrm{D}(X_1) \hookrightarrow \mathrm{D}(X_2)$. If $K_{X_2/X_1} = 0$ then F is an equivalence.

One might well imagine that a proof of Conjecture 2.1 would also apply to this much more general situation. See [26], [27] for more details on this conjecture.

One of the main problems with studying the effect of the minimal model programme on derived categories is the presence of singularities. Considering varieties with mild singularities is essential for the minimal model programme in dimension at least three. But many of the techniques that have been developed to understand derived categories rely on smoothness.

The situation in dimension three is rather special since there is an explicit list of local models for terminal singularities. Each is a finite quotient of a hypersurface singularity. The hypersurface singularity can be thought of as a special fibre of a smooth fourfold, which allows one to get around the bad behaviour of derived categories of singular varieties. In this way Chen [15] was able to extend Theorem 3.1 so as to allow Gorenstein terminal singularities.

It is easy to see that if some version of Conjecture 7.1 is to hold, then for certain singular varieties the derived category must be modified in some way. Thus for example there exist flops and flips which take one from a smooth variety to a singular one. But there can never be a fully faithful embedding $\mathrm{D}(Y_1) \hookrightarrow \mathrm{D}(Y_2)$ if Y_1 is singular and Y_2 smooth because of the following straightforward consequence of Serre's theorem on regularity of local rings.

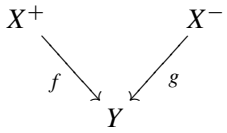
Proposition 7.2. *Let Y be a projective variety and $\mathrm{D}(Y)$ its bounded derived category of coherent sheaves. Then Y is smooth if and only if*

$$\dim_{\mathbb{C}} \bigoplus_{i\in\mathbb{Z}} \mathrm{Hom}^i_{\mathrm{D}(Y)}(A, B[i]) < \infty$$

for all objects A and B of $\mathrm{D}(Y)$.

At the moment this problem of how to modify the derived category for singular varieties is the biggest obstacle to progress in this area.

If a variety has only quotient singularities it is clear that one should consider the derived category of coherent sheaves on the corresponding stack. Kawamata explained this using the example of the Francia flip [25]. This is a threefold flip



in which X^- is smooth but X^+ has an isolated quotient singularity. By the argument above there cannot be an embedding $\mathrm{D}(X^+) \hookrightarrow \mathrm{D}(X^-)$. On the other hand there is an embedding $\mathrm{D}(\mathcal{X}^+) \hookrightarrow \mathrm{D}(X^-)$, where $\mathcal{X}^+ \to X^+$ is the Deligne–Mumford stack associated to the quotient singularity. In the same way, Chen's result on Gorenstein threefold flops extends to arbitrary terminal threefolds so long as one considers the derived categories of the corresponding stacks [26].

Kawamata has extended Conjecture 7.1 to a statement involving log varieties with quotient singularities. In an impressive piece of work [27] he was able to prove his conjecture under the additional assumption that the birational maps f_i were toroidal. Roughly speaking this means that they are locally defined by toric data. The derived McKay correspondence for abelian groups (Theorem 2.7) follows as a special case of this result.

Finally, Kawamata was able [28] to use his result together with the log minimal model programme for toric varieties to solve a long-standing problem: the existence of a full exceptional collection on any toric variety. For more on these state of the art developments we refer the reader to [27], [28].

References

[1] Abramovich, D., and Polishchuk, A., Sheaves of t-structures and valuative criteria for stable complexes. math.AG/0309435.

[2] Aspinwall, P., A point's point of view of stringy geometry. *J. High Energy Phys.* **1** (2003), 002.

[3] Beilinson, A., Bernstein, J., and Deligne, P., Faisceaux pervers. In *Analysis and topology on singular spaces, I* (Luminy, 1981), *Astérisque* **100** (1982), 5–171.

[4] Bezrukavnikov, R., and Kaledin, D., McKay equivalence for symplectic resolutions of quotient singularities. *Tr. Mat. Inst. Steklova* **246** (3) (2004), 20–42; English transl. *Proc. Steklov Inst. Math.* **246** (3) (2004), 13–33

[5] Bondal, A., and Orlov, D., Semiorthogonal decomposition for algebraic varieties. alg-geom/9506012.

[6] Bondal, A., and Orlov, D., Reconstruction of a variety from the derived category and groups of autoequivalences. *Compositio Math.* **125** (3) (2001), 327–344.

[7] Bondal, A., and Orlov, D., Derived categories of coherent sheaves. *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 47–56.

[8] Borcherds, R., Automorphic forms on $O_{s+2,2}(\mathbf{R})$ and infinite products. *Invent. Math.* **120** (1995), 161–213.

[9] Brenner, S., and Butler, M., Generalizations of the Bernstein-Gelfand-Ponomarev reflection functors. In *Representation theory, II* (Proc. Second Internat. Conf., Carleton Univ., Ottawa, Ont., 1979), Lecture Notes in Math. 832, Springer-Verlag, Berlin 1980, 103–169.

[10] Bridgeland, T., King, A., and Reid, M., The McKay correspondence as an equivalence of derived categories. *J. Amer. Math. Soc.* **14** (3) (2001), 535–554

[11] Bridgeland, T., Flops and derived categories. *Invent. Math.* **147** (3) (2002), 613–632.

[12] Bridgeland, T., Stability conditions on triangulated categories. *Ann. of Math.*, to appear; math.AG/0212237.

[13] Bridgeland, T., Stability conditions on K3 surfaces. math.AG/0307164.

[14] Bridgeland, T., Stability conditions and Kleinian singularities. math.AG/0508257.

[15] Chen, J.-C., Flops and equivalences of derived categories for threefolds with only terminal Gorenstein singularities. *J. Differential Geom.* **61** (2) (2002), 227–261.

[16] Craw, A., and Ishii, A., Flops of G-Hilb and equivalences of derived categories by variation of GIT quotient. *Duke Math. J.* **124** (2) (2004), 259–307.

[17] Douglas, M., Dirichlet branes, homological mirror symmetry, and stability. *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing 2002, 395–408.

[18] Happel, D., Reiten, I., and Smalo, S., *Tilting in abelian categories and quasitilted algebras*. Mem. Amer. Math. Soc. 120, Amer. Math. Soc., Providence, RI, 1996.

[19] Hille, L., and Van den Bergh, M., Fourier-Mukai transforms. math.AG/0402043.

[20] Horja, R., Derived category automorphisms from mirror symmetry. *Duke Math. J.* **127** (1) (2005), 1–34.

[21] Huybrechts, D., and Thomas, R., $\mathbb{P}$-objects and autoequivalences of derived categories. *Math. Res. Lett.* **13** (1) (2006), 87–98.

[22] Inaba, M., Toward a definition of moduli of complexes of coherent sheaves on a projective scheme. *J. Math. Kyoto Univ.* **42** (2) (2002), 317–329.

[23] Ishii, A., and Uehara, H., Autoequivalences of derived categories on the minimal resolutions of A_n-singularities on surfaces. *J. Differential Geom.* **71** (3) (2005), 385–435.

[24] Kawamata, Y., Crepant blowing-up of 3-dimensional canonical singularities and its application to degenerations of surfaces. *Ann. of Math.* **127** (1988), 93–163.

[25] Kawamata, Y., Francia's flip and derived categories. In *Algebraic geometry*, Walter de Gruyter, Berlin 2002, 197–215.

[26] Kawamata, Y., D-equivalence and K-equivalence. *J. Differential Geom.* **61** (1) (2002), 147–171.

[27] Kawamata, Y., Log crepant birational maps and derived categories. *J. Math. Sci. Univ. Tokyo* **12** (2) (2005), 211–231.

[28] Kawamata, Y., Derived Categories of Toric Varieties. math.AG/0503102.

[29] Kawamata, Y., Derived equivalence for stratified Mukai flop on $G(2,4)$. math.AG/0503101.

[30] Kollár, J., Flops. *Nagoya Math. J.* **113** (1989), 15–36.

[31] Kontsevich, M., Homological algebra of mirror symmetry. *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 120–139.

[32] Kuznetsov, A., Derived categories of cubic and V14 threefolds. *Tr. Mat. Inst. Steklova* **246** (2004), 183–207; English transl. *Proc. Steklov Inst. Math.* **246** (3) (2004), 171–194.

[33] Kuznetsov, A., Hyperplane sections and derived categories. math.AG/0503700.

[34] Kuznetsov, A., Homological Projective Duality. math.AG/0507292.

[35] Kuznetsov, A., Derived Categories of Quadric Fibrations and Intersections of Quadrics. math.AG/0510670.

[36] Macri, E., Some examples of moduli spaces of stability conditions on derived categories. math.AG/0411613.

[37] Mukai, S., Duality between $D(X)$ and $D(\hat{X})$ with its application to Picard sheaves. *Nagoya Math. J.* **81** (1981), 153–175.

[38] Mukai, S., On the moduli space of bundles on K3 surfaces I. In *Vector Bundles on Algebraic Varieties* (ed. by M. F. Atiyah et al.), Tata Inst. Fund. Res. Stud. Math. 11, Oxford University Press, New York 1987, 341–413.

[39] Namikawa, Y., Mukai flops and derived categories. *J. Reine Angew. Math.* **560** (2003), 65–76.

[40] Namikawa, Y., Mukai flops and derived categories. II. In *Algebraic structures and moduli spaces*, CRM Proc. Lecture Notes 38, Amer. Math. Soc., Providence, RI, 2004, 149–175.

[41] Okada, S., Stability manifold of $\mathbb{P}^1$. math.AG/0411220.

[42] Orlov, D., Derived categories of coherent sheaves on abelian varieties and equivalences between them. *Izv. Ross. Akad. Nauk Ser. Mat.* **66** (3) (2002), 131–158; English transl. *Izv. Math.* **66** (3) (2002), 569–594.

[43] Reid, M., McKay correspondence. alg-geom/9702016.

[44] Reid, M., La correspondance de McKay. Séminaire Bourbaki 1999/2000, *Astérisque* **276** (2002), 53–72.

[45] Rouquier, R., Categories derivees et geometrie birationnelle. math.AG/0503548.

[46] Seidel, P., and Thomas, R. P., Braid group actions on derived categories of coherent sheaves. *Duke Math. J.* **108** (1) (2001), 37–108, math.AG/0001043.

[47] Szendröi, B., Artin group actions on derived categories of threefolds. *J. Reine Angew. Math.* **572** (2004), 139–166.

[48] Toda, Y., Stability conditions and crepant small resolutions, preprint.

[49] Toen, B., and Vaquie, M., Moduli of objects in dg-categories. math.AG/0503269.

[50] Van den Bergh, M., Three-dimensional flops and noncommutative rings. *Duke Math. J.* **122** (3) (2004), 423–455.

[51] Wierzba, J., and Wisniewski, J. A., Small contractions of symplectic 4-folds. math.AG/0201028

Department of Pure Mathematics, University of Sheffield, Hicks Building, Hounslow Road, Sheffield, S3 7RH, UK

E-mail: t.bridgeland@sheffield.ac.uk

Invariants of singularities of pairs

Lawrence Ein and Mircea Mustaţă*

Abstract. Let X be a smooth complex variety and Y be a closed subvariety of X, or more generally, a closed subscheme of X. We are interested in invariants attached to the singularities of the pair (X, Y). We discuss various methods to construct such invariants, coming from the theory of multiplier ideals, D-modules, the geometry of the space of arcs and characteristic p techniques. We present several applications of these invariants to algebra, higher dimensional birational geometry and to singularities.

Mathematics Subject Classification (2000). 14J17, 14J40, 14E05, 14E30.

Keywords. Singularities, birational maps.

1. Introduction

Let X be a smooth complex variety of dimension n and Y be a closed subvariety of X (or more generally, a closed subscheme of X). We are interested in studying the singularities of the pair (X, Y). The general setup is to assume only that X is normal and $\mathbb{Q}$-Gorenstein, as in [32]. However, several of the approaches we will discuss become particularly transparent if we assume, as we do, the smoothness of the ambient variety. The following are some examples of pairs.

Examples 1.1. (i) Let $X = \mathbb{C}^n$ and let Y be a hypersurface defined by an equation $f(x_1, \ldots, x_n) = 0$. For instance f can be the Fermat hypersurface $x_1^m + x_2^m + \cdots + x_n^m = 0$, which has an isolated singularity of multiplicity m at the origin.

(ii) If X is a smooth projective variety and L is a line bundle on X, then we take Y to be the base locus of the complete linear system $|L|$, i.e. $Y = \bigcap_{D \in |L|} D$.

(iii) Let X be a smooth affine variety with coordinate ring R. If $I \subseteq R$ is an ideal, then we take Y to be the closed subscheme defined by I.

In what follows we present various invariants attached to such pairs and we discuss some of their applications. Our main point is that the same invariants that play an important role in higher dimensional algebraic geometry arise also in several other approaches to singularities.

*Research of Ein and Mustaţă was partially supported by the NSF under grants DMS 0200278 and 0500127.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. Multiplier ideals

Multiplier ideals were first introduced by J. Kohn for solving certain partial differential equations. Siu and Nadel introduced them to complex geometry. We discuss below these ideals in the context of algebraic geometry.

Let X be a smooth complex affine variety and Y be a closed subscheme of X. Suppose that the ideal of Y is generated by $f_1, \ldots, f_m$, and let λ be a positive real number. We define the multiplier ideal of (X, Y) of coefficient λ as follows:

$$\mathcal{J}(X, \lambda \cdot Y) = \left\{ g \in \mathcal{O}_X \mid \frac{|g|^2}{\left(\sum_{i=1}^m |f_i|^2\right)^{\lambda}} \text{ is locally integrable} \right\}.$$

Example 2.1. Let $X = \mathbb{C}^n$ and let Y be the closed subscheme of X defined by $f = x_1^{a_1} \ldots x_n^{a_n}$. Then

$$\mathcal{J}(X, \lambda \cdot Y) = (x_1^{\lfloor \lambda a_1 \rfloor} \ldots x_n^{\lfloor \lambda a_n \rfloor}),$$

where $\lfloor \alpha \rfloor$ denotes the integer part of α. Equivalently, if H_i is the divisor defined by $x_i = 0$, then g is in $\mathcal{J}(X, \lambda \cdot Y)$ if and only if

$$\operatorname{ord}_{H_i} g \geq \lfloor \lambda a_i \rfloor,$$

for $i = 1, \ldots, n$.

We can use a log resolution of singularities and the above example to give in general a more geometric description of the multiplier ideals of (X, Y). By Hironaka's Theorem there is a log resolution of singularities of the pair (X, Y), i.e. a proper birational morphism

$$\mu \colon X' \longrightarrow X$$

with the following properties. The variety X' is smooth, $\mu^{-1}(Y)$ is a divisor, and the union of $\mu^{-1}(Y)$ and the exceptional locus of μ has simple normal crossings. The relative canonical divisor $K_{X'/X}$ is locally defined by the determinant of the Jacobian $J(\mu)$ of μ, hence it is supported on the exceptional locus of μ. We write $\mu^{-1}(Y) = \sum_{i=1}^N a_i E_i$ and $K_{X'/X} = \sum_{i=1}^N k_i E_i$, where the E_i are distinct smooth irreducible divisors in X' such that $\sum_{i=1}^N E_i$ has only simple normal crossing singularities.

Suppose that $x_1, \ldots, x_n$ are local coordinates in X and $y_1, y_2, \ldots, y_n$ are local coordinates for an open set in X'. Note that

$$\mu^* dx_1 \ldots dx_n d\bar{x}_1 \ldots d\bar{x}_n = |\det(J(\mu)|^2 dy_1 \ldots dy_n d\bar{y}_1 \ldots d\bar{y}_n. \tag{1}$$

The local integrability of a function g on X can be expressed as a local integrability condition on X' via the change of variable formula. This reduces us to a monomial situation, similar to that in Example 2.1. On deduces that $g \in \mathcal{J}(X, \lambda \cdot Y)$ if and only if

$$\operatorname{ord}_{E_i} g \geq \lfloor \lambda a_i \rfloor - k_i$$

for every i. Equivalently, if we put $\lfloor \lambda \mu^{-1}(Y) \rfloor = \sum_i \lfloor \lambda a_i E_i \rfloor$, then

$$\mathcal{J}(X, \lambda \cdot Y) = \mu_* \mathcal{O}_{X'}(K_{X'/X} - \lfloor \lambda \mu^{-1}(Y) \rfloor). \tag{2}$$

We refer to [34] for details.

Note that because of the original definition, it follows that this expression for $\mathcal{J}(X, \lambda \cdot Y)$ is independent of the choice of a resolution of singularities. On the other hand, the formula (2) applies also when X is non necessarily affine. Note also that this formula implies that if $\lambda_1 \geq \lambda_2$, then

$$\mathcal{J}(X, \lambda_1 \cdot Y) \subseteq \mathcal{J}(X, \lambda_2 \cdot Y).$$

If λ is small enough, then $\lambda a_i < k_i + 1$ for $i = 1, \ldots, N$. This implies that

$$\operatorname{ord}_{E_i} 1 \geq \lfloor \lambda a_i \rfloor - k_i,$$

hence $\mathcal{J}(X, \lambda \cdot Y) = \mathcal{O}_X$. This leads us to the definition of the log canonical threshold of the pair (X, Y): this is the smallest λ such that $\mathcal{J}(X, \lambda \cdot Y) \neq \mathcal{O}_X$, i.e.

$$c = \operatorname{lc}(X, Y) = \min_i \left\{ \frac{k_i + 1}{a_i} \right\}.$$

We may regard $\frac{1}{c}$ as a refined version of multiplicity. In general a singularity with a smaller log canonical threshold tends to be more complex.

The first appearance of the log canonical threshold was in the work of Arnold, Gusein-Zade and Varchenko (see [2] and [48]), in connection with the behavior of certain integrals over vanishing cycles. In the last decade this invariant has enjoyed renewed interest due to its applications to birational geometry. The following is probably the most interesting open problem about log canonical thresholds.

Conjecture 2.2 (Shokurov). For every n, the set

$$\{\operatorname{lc}(X, Y) \mid \dim(X) = n, Y \subset X\}$$

satisfies the Ascending Chain Condition: it contains no strictly increasing sequences.

We can consider also higher jumping numbers. In general, we say that λ is a jumping number of (X, Y), if

$$\mathcal{J}(X, \lambda \cdot Y) \subsetneq \mathcal{J}(X, (\lambda - \varepsilon) \cdot Y)$$

for all $\varepsilon > 0$. If λa_i is not an integer, then $\lfloor \lambda a_i \rfloor = \lfloor (\lambda - \varepsilon) a_i \rfloor$ for sufficiently small positive ε. We see that a necessary condition for λ to be a jumping number is that λa_i is an integer for some i. In particular, if λ is a jumping number, then it is rational and has a bounded denominator.

The following theorem gives a periodicity property of the jumping numbers.

Theorem 2.3. (i) *If $Y = D$ is a hypersurface in X, then*

$$\mathcal{J}(X, \lambda \cdot D) \cdot \mathcal{O}_X(-D) = \mathcal{J}(X, (\lambda + 1) \cdot D).$$

(ii) (Ein and Lazarsfeld [16]) *For every Y defined by the ideal I_Y, if $\lambda \geq \dim X - 1$, then*

$$\mathcal{J}(X, \lambda \cdot Y) \cdot I_Y = \mathcal{J}(X, (\lambda + 1) \cdot Y).$$

Corollary 2.4. *If $\lambda > \dim X - 1$, then λ is a jumping number for (X, Y) if and only if so is $(\lambda + 1)$.*

We conclude that the set of jumping numbers of the pair (X, Y) is a discrete subset of $\mathbb{Q}$ and it is eventually periodic with period one.

Example 2.5. If Y is a smooth subvariety of X of codimension e, then the set of jumping numbers of the pair (X, Y) is $\{e, e + 1, \dots\}$. In particular $\mathrm{lc}(X, Y) = e$.

Example 2.6. (Howald) Let $X = \mathbb{C}^n$ and let Y be the closed subscheme defined by a monomial ideal $\mathfrak{a}$. If $a = (a_1, a_2, \dots, a_n) \in \mathbb{N}^n$, we denote the monomial $x_1^{a_1} \dots x_n^{a_n}$ by x^a. Consider the Newton polyhedron $P_{\mathfrak{a}}$ associated with $\mathfrak{a}$: this is the convex hull of those $a \in \mathbb{N}^n$ such that $x^a \in \mathfrak{a}$. Using toric geometry Howald showed in [27] that

$$\mathcal{J}(X, \lambda \cdot Y) = (x^a \mid a + e \in \mathrm{Int}(\lambda \cdot P_{\mathfrak{a}})),$$

where $e = (1, \dots, 1)$. In particular, the log canonical threshold c of (X, Y) is characterized by the fact $c \cdot e$ lies on the boundary of $P_{\mathfrak{a}}$.

For example, suppose that $\mathfrak{a}$ is the ideal $(x_1^{a_1}, \dots, x_n^{a_n})$. In this case, the boundary of $P_{\mathfrak{a}}$ is

$$\Big\{u = (u_1, \dots, u_n) \in \mathbb{R}_+^n \mid \textstyle\sum_{i=1}^n \frac{u_i}{a_i} = 1\Big\}.$$

Therefore $\mathrm{lc}(X, Y) = \sum_i \frac{1}{a_i}$.

Example 2.7. Suppose that $X = \mathbb{C}^2$ and Y is the plane cuspidal curve defined by $x^3 + y^5 = 0$. Then the set of jumping numbers for Y is periodic with period 1. The jumping numbers in $(0, 1]$ are $\big\{\frac{8}{15}, \frac{11}{15}, \frac{13}{15}, \frac{14}{15}, 1\big\}$.

One reason that multiplier ideals have been very powerful in studying questions in higher dimensional algebraic geometry is that they appear naturally in a Kodaira type vanishing theorem. The following statement is the algebraic version of a result due to Nadel. In our context, it can be deduced from the Kawamata–Viehweg Vanishing Theorem (see [34]).

Theorem 2.8. *Let X be a smooth projective variety and Y a closed subscheme of X defined by the ideal I_Y. If A is a line bundle such that $I_Y \otimes A$ is globally generated, and if L is a line bundle such that $L - A$ is big and nef, then for every $i > 0$*

$$H^i(X, \mathcal{O}_X(K_X + L) \otimes \mathcal{J}(X, Y)) = 0.$$

3. Applications of multiplier ideals

One of the most important applications of multiplier ideals is the following theorem of Siu (see [44] and [45]) on the deformation invariance of plurigenera.

Theorem 3.1. *Let $f : X \longrightarrow T$ be a smooth projective morphism of relative dimension n between two smooth irreducible varieties. If we denote by X_t the fiber $f^{-1}(t)$ for each $t \in T$, then for every fixed $m > 0$, the dimension of the cohomology group $H^0(X_t, (\Omega^n_{X_t})^{\otimes m})$ is independent of t.*

The techniques involved in the proof of this theorem have been recently applied by Siu, Hacon and McKernan to study one of the outstanding problems in higher dimensional algebraic geometry, the finite generation of the canonical ring (see, for example, [23]).

In a different direction, there are applications of multiplier ideals to singularities of theta divisors on abelian varieties. Let (X, Θ) be a principally polarized abelian variety, that is, Θ is an ample divisor on an abelian variety X such that $\dim H^0(X, \mathcal{O}_X(\Theta)) = 1$. The following result is due to Ein and Lazarsfeld [17].

Theorem 3.2. *Let (X, Θ) be a principally polarized abelian variety. If Θ is irreducible, then Θ has at most rational singularities.*

Corollary 3.3. *Let (X, Θ) be a principally polarized abelian variety of dimension g, with Θ irreducible. If*

$$\Sigma_k(\Theta) = \{x \in X \mid \mathrm{mult}_x(\Theta) \geq k\},$$

then for every $k \geq 2$ we have $\mathrm{codim}(\Sigma_k(\Theta), X) \geq k+1$. In particular, Θ is a normal variety and $\mathrm{mult}_x(\Theta) \leq g - 1$ for every singular point x on Θ.

Remark 3.4. The fact that Θ is normal was first conjectured by Arbarello, De Concini and Beauville. When X is the Jacobian of a curve, the fact that Θ has only rational singularities was proved by Kempf. Note also that in this case, a classical theorem of Riemann expresses the multiplicity of Θ at a point in term of the dimension of the corresponding linear system on the curve. It was Kollár who first observed in [31] that one can use vanishing theorems to study the singularities of the theta divisor: he showed that for every principally polarized abelian variety (X, Θ), we have $\mathrm{lc}(X, \Theta) = 1$. Theorem 3.2 above is a strengthening of Kollár's result.

Multiplier ideals have been applied in several other directions: to Fujita's problem on adjoint linear systems [3], to Effective Nullstellensatz [16], to Effective Artin–Rees Theorem [20]. Building on work of Tsuji, recently Hacon and McKernan and independently, Takayama have used multiplier ideals to prove a very interesting result on boundedness of pluricanonical maps for varieties of general type (see [24] and [47]). We end this section with an application to commutative algebra due to Ein, Lazarsfeld and Smith [19].

Let X be a smooth n-dimensional variety and $Y \subseteq X$ defined by the reduced sheaf of ideals $\mathfrak{a}$. The m^{th} symbolic power of $\mathfrak{a}$ is the sheaf $\mathfrak{a}^{(m)}$ of functions on X that vanish with multiplicity at least m at the generic point of every irreducible component of Y. If Y is smooth, then the symbolic powers of $\mathfrak{a}$ agree with the usual powers, but in general they are very different.

Theorem 3.5. *If X is a smooth n-dimensional variety and if $\mathfrak{a}$ is a reduced sheaf of ideals, then $\mathfrak{a}^{(mn)} \subseteq \mathfrak{a}^m$ for every m.*

4. Bounds on log canonical thresholds and birational rigidity

In this section we compare the log canonical threshold with the classical Samuel multiplicity. We give then an application of the inequality between these two invariants to a classical question on birational rigidity. Let X be a smooth complex variety and $x \in X$ a point. Denote by R the local ring of X at x, and by $\mathfrak{m}$ its maximal ideal. The following result was proved by de Fernex, Ein and Mustaţă in [10].

Theorem 4.1. *Let $\mathfrak{a}$ be an ideal in R that defines a subscheme Y supported at x. Let c be the log canonical threshold of (X, Y), $l(R/\mathfrak{a})$ be the length of $R/\mathfrak{a}$ and $e(\mathfrak{a})$ be the Samuel multiplicity of R along $\mathfrak{a}$. If $n = \dim R$, then we have the following inequalities.*

(i) $l(R/\mathfrak{a}) \geq \frac{n^n}{n! \cdot c^n}$.

(ii) $e(\mathfrak{a}) \geq \frac{n^n}{c^n}$. *Furthermore, this is an equality if and only if the integral closure of $\mathfrak{a}$ is equal to $\mathfrak{m}^k$ for some k.*

The first assertion in (ii) above can be easily deduced from (i). The proof of (i) proceeds by reduction to the monomial case, via a Gröbner deformation. When $\mathfrak{a}$ is monomial, the inequality follows by a combinatorial argument from the explicit description of the invariants.

Example 4.2. Suppose that $\mathfrak{a} = (x_1^{a_1}, \ldots, x_n^{a_n})$. In this case $e(\mathfrak{a}) = \prod_{i=1}^n a_i$ and $\mathrm{lc}(\mathfrak{a}) = \sum_{i=1}^n \frac{1}{a_i}$. The inequality in Theorem 4.1(ii) becomes

$$\prod_{i=1}^n a_i \geq \frac{n^n}{\left(\sum_{i=1}^n \frac{1}{a_i}\right)^n}.$$

This is equivalent to

$$\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{a_i}\right)^n \geq \prod_{i=1}^n \frac{1}{a_i},$$

which is just the classical inequality between the arithmetic and the geometric mean.

Remark 4.3. When X is a surface, the inequality in (ii) above was first proved by Corti [9].

Theorem 4.1 is used in [11] to study the behavior of the log canonical threshold under a generic projection. More generally, one proves the following

Theorem 4.4. *Let $f : X \longrightarrow Y$ be a smooth proper morphism of relative dimension k between two smooth complex varieties. If V is a locally complete intersection of codimension k in X such that $f|_V$ is finite, then*

$$\mathrm{lc}(Y,\, f_*(V)) \leq \frac{\mathrm{lc}(X,\, V)^k}{k^k}.$$

Using Theorem 4.4 and some beautiful geometric ideas of Pukhlikov [42], one gives in [11] a simple uniform proof for the following result.

Theorem 4.5. *If X is a smooth hypersurface of degree N in $\mathbb{CP}^N$, with $4 \leq N \leq 12$, then X is birationally superrigid. In particular, every birational automorphism of X is biregular.*

Remark 4.6. Consider the group $\mathrm{Aut}_{\mathbb{C}}(\mathbb{C}(X))$, the automorphism group of the field of rational functions of X. This is naturally isomorphic to $\mathrm{Bir}_{\mathbb{C}}(X)$, the group of birational automorphisms of X. If X is birationally superrigid, then $\mathrm{Bir}_{\mathbb{C}}(X) \simeq \mathrm{Aut}_{\mathbb{C}}(X)$, the automorphism group of X. When X is a hypersurface of degree N in $\mathbb{P}^N$, X has no nonzero vector fields and therefore $\mathrm{Aut}_{\mathbb{C}}(X)$ is a finite group. This shows that X is not a rational variety: if $\mathbb{C}(X)$ is purely transcendental, then $\mathrm{Aut}_{\mathbb{C}}(X)$ will contain a subgroup isomorphic to the general linear group GL_n.

Remark 4.7. When $N = 4$, it is a classical theorem of Iskovskikh and Manin that X is birationally rigid [28]. They used this to show that the function field of a suitable quartic threefold provides a counterexample to the classical Lüroth problem: $\mathbb{C}(X)$ is a $\mathbb{C}$-subfield of the purely transcendental field $\mathbb{C}(x_1, x_2, x_3)$ but $\mathbb{C}(X)$ is not purely transcendental. For $N = 5$, Theorem 4.5 is a result of Pukhlikov [40]. The cases $N = 6$, 7 and 8 were first proved by Cheltsov [8]. We mention also that Pukhlikov [41] has shown that a generic hypersurface of degree N in $\mathbb{P}^N$ is superrigid for every $N \geq 4$.

5. Bernstein–Sato polynomials

Let $f \in \mathbb{C}[x_1, x_2, \ldots, x_n]$ be a nonzero polynomial. We denote by A_n the Weyl algebra of differential operators on $\mathbb{A}^n$, that is

$$A_n = \mathbb{C}[x_1, \ldots, x_n, \partial_{x_1}, \ldots, \partial_{x_n}].$$

Let s be another variable and consider the following functional equation:

$$b(s) f^s = P(s, x, \partial_x) \bullet f^{s+1}, \tag{3}$$

where $b(s) \in \mathbb{C}[s]$ and $P \in A_n[s]$. Here f^s is considered a formal symbol, and the action $\bullet$ of P is defined via $\partial_{x_i} \bullet f^s = sf^{-1}\frac{\partial f}{\partial x_i} f^s$. On the other hand, if we let $s = m$ for an integer m, then (3) has the obvious meaning.

It is easy to see that the set of polynomials $b(s)$ for which there is P satisfying (3) is an ideal in the polynomial ring $\mathbb{C}[s]$. It was proved by Bernstein in [4] using the theory of holonomic D-modules that this ideal is nonzero. Its monic generator is denoted by $b_f(s)$ and is called the Bernstein–Sato polynomial of f. It is an interesting and subtle invariant of the singularities of the hypersurface defined by f.

Examples 5.1. (1) Making $s = 1$ in (3) we see that $b_f(-1)f^{-1}$ lies in $\mathbb{C}[x_1, \ldots, x_n]$. If f is nonconstant, it follows that -1 is a root of b_f.

(2) If $f = x$, then $b_f(s) = (s + 1)$. Indeed, we have

$$(s + 1)f^s = \partial_x \bullet f^{s+1}.$$

More generally, if f defines a nonsingular hypersurface, then $b_f(s) = (s + 1)$.

(3) If $f = x_1^2 + \cdots + x_n^2$, then $b_f(s) = (s + 1)\left(s + \frac{n}{2}\right)$ and

$$b_f(s)f^s = \frac{1}{4}(\partial_{x_1}^2 + \cdots + \partial_{x_n}^2) \bullet f^{s+1}.$$

(4) If $f = x^2 + y^3$, then $b_f(s) = \left(s + \frac{5}{6}\right)(s + 1)\left(s + \frac{7}{6}\right)$ and

$$b_f(s)f^s = \left(\frac{1}{27}\partial_y^3 + \frac{1}{6}y\partial_x^2\partial_y + \frac{1}{8}x\partial_x^3 + \frac{3}{8}\partial_x^2\right) \bullet f^{s+1}.$$

Computing Bernstein–Sato polynomials in general is quite subtle (see [49]). On the other hand, there has been a lot of recent progress in algorithmic computation using Gröbner bases in the Weyl algebra (see [43]).

We describe now the connection between the roots of the Bernstein–Sato polynomial of f and the jumping numbers of the hypersurface Y defined in $\mathbb{A}^n$ by f. An important theorem of Kashiwara [30] asserts that all the roots of $b_f(s)$ are negative rational numbers. Building on Kashiwara's work, Lichtin made this more explicit in [35], describing a connection between the roots of $b_f(s)$ and a log resolution of the pair $(\mathbb{A}^n, Y)$. This says that if $\mu\colon X' \to \mathbb{A}^n$ is a log resolution of (X, Y), then every root of $b_f(s)$ is of the form $-\frac{k_i+m}{a_i}$ for some i and some positive integer m (we use the notation introduced in §2). In particular, we see that every root of $b_f(s)$ is rational, and no larger than $-\mathrm{lc}(\mathbb{A}^n, Y)$. However, we stress that unlike in the case of multiplier ideals, there is no explicit description of the Bernstein–Sato polynomial in terms of a log resolution.

On the other hand, the following result of Ein, Lazarsfeld, Smith and Varolin [20] shows that in a suitable range, all jumping numbers give roots of the Bernstein–Sato polynomial.

Theorem 5.2. *If $\lambda \in (0, 1]$ is a jumping number of $(\mathbb{A}^n, Y)$, then $-\lambda$ is a root of the Bernstein–Sato polynomial $b_f(s)$.*

The proof of this theorem uses the functional equation (3) and integration by parts. The case when $\lambda = \mathrm{lc}(\mathbb{A}^n, Y)$ was proved by Kollár in [32]. Note that in conjunction with the above mentioned result of Lichtin, this gives the following

Corollary 5.3. *The largest root of $b_f(s)$ is $-\mathrm{lc}(\mathbb{A}^n, Y)$.*

A different point of view on the connection between multiplier ideals and Bernstein–Sato polynomials was given by Budur and Saito. In fact, they show how to recover the multiplier ideals from a filtration that appears in D-module theory, the V-filtration. We present now their result.

Let t be a new variable, and let A_{n+1} denote the Weyl algebra corresponding to the affine space $\mathbb{A}^{n+1}$, with coordinates $x_1, \ldots, x_n, t$. We consider the module B_f that is the first local cohomology module of $\mathbb{A}^{n+1}$ along the embedding of $\mathbb{A}^n$ as the graph of f, i.e.

$$B_f = \mathbb{C}[x_1, \ldots, x_n, t]_{f-t}/\mathbb{C}[x_1, \ldots, x_n, t].$$

Let δ be the class of $\frac{1}{f-t}$ in B_f (δ is the "delta-function associated to the graph of f").

Note that B_f has a natural structure of left module over A_{n+1}. Since $\partial_t^m \delta$ is the class of $\frac{m!}{(f-t)^{m+1}}$ in B_f, we see that B_f is free over $\mathbb{C}[x_1, \ldots, x_n]$, with basis given by $\{\partial_t^j \delta \mid j \geq 0\}$.

The V-filtration is a decreasing filtration on B_f by left A_n-submodules V^α indexed by $\alpha \in \mathbb{Q}$, with the following properties:

(i) $\bigcup_\alpha V^\alpha = B_f$.

(ii) The filtration is semicontinuous and discrete in the following sense: there is a positive integer ℓ such that for every integer m and every $\alpha \in \left(\frac{m-1}{\ell}, \frac{m}{\ell}\right]$ we have $V^\alpha = V^{m/\ell}$.

(iii) We have $t \cdot V^\alpha \subseteq V^{\alpha+1}$ for every α, with equality if $\alpha > 0$.

(iv) We have $\partial_t \cdot V^\alpha \subseteq V^{\alpha-1}$ for every α.

(v) For every α, if we put $V^{>\alpha} := \cup_{\beta>\alpha} V^\beta$, then $(\partial_t t - \alpha)$ is nilpotent on $V^\alpha/V^{>\alpha}$.

The key property is (v) above. One can think of the V-filtration as an attempt to diagonalize the operator $\partial_t t$ on B_f. It is not hard to show that if a filtration as above exists, then it is unique. Malgrange [36] proved the existence of the V-filtration using the existence of the Bernstein–Sato polynomial and the rationality of its roots. To explain the role played by $b_f(s)$ in the construction of the V-filtration, we mention that the equation (3) in the definition of b_f is equivalent with the following equality in B_f:

$$b(-\partial_t t) \cdot \delta = P(-\partial_t t, x, \partial_x) \cdot t\delta. \tag{4}$$

The following result of Budur and Saito [5] shows that the multiplier ideals can be obtained as a piece of the V-filtration. We consider $\mathbb{C}[x_1, \ldots, x_n]$ embedded in B_f by $h \to h\delta$.

Theorem 5.4. *If Y is the hypersurface defined by f, then for every $\lambda > 0$ we have $\mathcal{J}(\mathbb{A}^n, \lambda \cdot Y) = V^{\lambda+\varepsilon} \cap \mathbb{C}[x_1, \ldots, x_n]$, where $0 < \varepsilon \ll 1$.*

The assertion in Theorem 5.2 can be deduced from this statement. The proof of Theorem 5.4 involves two steps. First, one describes the V-filtration in the case when f defines a divisor with simple normal crossings: $f = x_1^{a_1} \ldots x_n^{a_n}$. In this case, let us put $\mathcal{J}'(\mathbb{A}^n, \alpha \cdot Y) := \mathcal{J}(\mathbb{A}^n, (\alpha - \varepsilon) \cdot Y)$ for $0 < \varepsilon \ll 1$ (with the convention $\mathcal{J}'(\mathbb{A}^n, \alpha \cdot Y) = \mathbb{C}[x_1, \ldots, x_n]$ if $\alpha \leq 0$). If we take V^α to be generated over A_n by $\mathcal{J}'(\mathbb{A}^n, (\alpha + j) \cdot Y)\partial_t^j \delta$, where j varies over the nonnegative integers, then one can check that these V^α satisfy the properties in the definition of the V-filtration. In particular, this easily implies the statement of Theorem 5.4 in this case. The hard part of the proof uses Saito's theory of mixed Hodge modules to deduce the general case of the theorem by relating the V-filtrations of f and of a log resolution.

We mention that Kashiwara constructed in [29] a V-filtration associated to several polynomials. Budur, Mustaţă and Saito used this in [7] to introduce and study the Bernstein–Sato polynomial associated to a subscheme not necessarily of codimension one, and to generalize Theorems 5.2 and 5.4 to this setting.

6. Spaces of arcs and contact loci

Let X be a smooth n-dimensional complex variety. Given $m \geq 0$, we denote by

$$X_m = \mathrm{Hom}(\mathrm{Spec}\, \mathbb{C}[t]/(t^{m+1}), X)$$

the space of m^{th} order jets on X. This carries a natural scheme structure. Similarly we define the space of formal arcs on X as

$$X_\infty = \mathrm{Hom}(\mathrm{Spec}\, \mathbb{C}[[t]], X).$$

These constructions are functorial, hence to every morphism $\mu\colon X' \to X$ we associate corresponding morphisms μ_m and μ_∞. Thanks to the work of Kontsevich, Denef, Loeser and others on motivic integration, in recent years these spaces have been very useful in constructing invariants of singular algebraic varieties. In what follows we describe some applications of these spaces to the study of singularities.

We have natural projection maps induced by truncation $X_{m+1} \to X_m$. Since X is smooth, this is locally trivial in the Zariski topology, with fiber $\mathbb{A}^n$. We similarly have projection maps $X_\infty \to X_m$. A subset C of X_∞ is called a *cylinder* if it is the inverse image of a constructible set S in some X_m. Moreover, C is called locally closed (closed, irreducible) if S is so. If C is a closed cylinder that is the inverse image of $S \subset X_m$, its codimension in X_∞ is equal to the codimension of S in X_m.

Consider a nonzero ideal sheaf $\mathfrak{a} \subseteq \mathcal{O}_X$ defining a subscheme $Y \subset X$. Given a finite jet or an arc γ on X, the order of vanishing of $\mathfrak{a}$ – or the order of contact of the

corresponding scheme Y – along γ is defined in the natural way. Specifically, pulling $\mathfrak{a}$ back via γ yields an ideal (t^e) in $\mathbb{C}[t]/(t^{m+1})$ or $\mathbb{C}[[t]]$, and one sets

$$\mathrm{ord}_\gamma(\mathfrak{a}) = \mathrm{ord}_\gamma(Y) = e.$$

(Take $\mathrm{ord}_\gamma(\mathfrak{a}) = m+1$ when $\mathfrak{a}$ pulls back to the zero ideal in $\mathbb{C}[t]/(t^{m+1})$ and $\mathrm{ord}_\gamma(\mathfrak{a}) = \infty$ when it pulls back to the zero ideal in $\mathbb{C}[[t]]$.) For a fixed integer $p \geq 0$, we define the *contact locus*

$$\mathrm{Cont}^p(Y) = \mathrm{Cont}^p(\mathfrak{a}) =_{\mathrm{def}} \{\gamma \in X_\infty \mid \mathrm{ord}_\gamma(\mathfrak{a}) = p\}.$$

Note that this is a locally closed cylinder: for $m \geq p$, it is the inverse image of

$$\mathrm{Cont}^p(Y)_m = \mathrm{Cont}^p(\mathfrak{a})_m =_{\mathrm{def}} \{\gamma \in X_m \mid \mathrm{ord}_\gamma(\mathfrak{a}) = p\}, \tag{5}$$

which is locally closed in X_m. A subset of X_∞ is called an *irreducible closed contact subvariety* if it is the closure of an irreducible component of $\mathrm{Cont}^p(Y)$ for some p and Y.

Suppose now that W is an arbitrary irreducible closed cylinder in X_∞. We can naturally associate a valuation of the function field of X to W as follows. If f is a nonzero rational function of X, we put

$$\mathrm{val}_W(f) = \mathrm{ord}_\gamma(f) \quad \text{for a general } \gamma \in W.$$

This valuation is not identically zero if and only if W does not dominate X.

If $\mu\colon X' \longrightarrow X$ is a proper birational morphism, with X' smooth, and if E is an irreducible divisor on X', then we define a valuation by

$$\mathrm{val}_E(f) = \text{ the vanishing order of } f \circ \mu \text{ along } E.$$

A valuation on the function field of X is called a *divisorial valuation* (with center on X) if it is of the form $m \cdot \mathrm{val}_E$ for some positive integer m and some divisor E as above.

A key invariant associated to a divisorial valuation v is its *log discrepancy*. If E is a divisor as above, we put $k_E = \mathrm{val}_E(\det(J(\mu))$, where $J(\mu)$ is the Jacobian matrix of μ. Equivalently, k_E is the coefficient of E in the relative canonical divisor $K_{X'/X}$. Note that k_E depends only on val_E (it does not depend on the model X'). Given an arbitrary divisorial valuation $m \cdot \mathrm{val}_E$, we define its log discrepancy as $m(k_E + 1)$.

Consider a divisor E on X' as above. If $C_m(E)$ is the closure of $\mu_\infty(\mathrm{Cont}^m(E))$, then it is not hard to see that $C_m(E)$ is an irreducible closed contact subvariety of X_∞ such that $\mathrm{val}_{C_m(E)} = m \cdot \mathrm{val}_E$. The following result of Ein, Lazarsfeld and Mustaţă [18] describes in general the connection between cylinders and divisorial valuations.

Theorem 6.1. *Let X be a smooth variety.*

(i) *If W is an irreducible, closed cylinder in X_∞ that does not dominate X, then the valuation* val_W *is divisorial.*

(ii) *For every divisorial valuation* $m \cdot \mathrm{val}_E$*, there is a unique maximal irreducible closed cylinder* W *such that* $\mathrm{val}_W = m \cdot \mathrm{val}_E$*: this is* $W = C_m(E)$.

(iii) *The map that sends* $m \cdot \mathrm{val}_E$ *to* $C_m(E)$ *gives a bijection between divisorial valuations of* $\mathbb{C}(X)$ *with center on* X *and the set of irreducible closed contact subvarieties of* X_∞.

The applicability of this result to the study of singularities is due to the following description of log discrepancy of a divisorial valuation in terms of the codimension of a certain set of arcs.

Theorem 6.2. *Given a divisorial valuation* $v = m \cdot \mathrm{val}_E$ *with center on* X*, if* $C_m(E)$ *is its associated irreducible closed contact subvariety in* X_∞*, then the log discrepancy of* v *is equal to* $\mathrm{codim}(C_m(E), X_\infty)$.

Combining the statements of the above theorems, we deduce a lower bound for the codimension of an arbitrary cylinder in terms of the log discrepancy of the corresponding divisor.

Corollary 6.3. *If* W *is a closed, irreducible cylinder in* X_∞ *that does not dominate* X*, then* $\mathrm{codim}(W, X_\infty)$ *is bounded below by the log discrepancy of* val_W.

Remark 6.4. The above two theorems also hold for singular varieties after some minor modifications using Nash's blow-up and Mather's canonical class.

The key ingredient in the proof of the above theorems is the following result due to Kontsevich, Denef and Loeser (see [14]. It constitutes the geometric content of the so-called Change of Variable Theorem in motivic integration. Suppose that $\mu\colon X' \longrightarrow X$ is a proper, birational morphism of smooth varieties and let $K_{X'/X}$ be the relative canonical divisor.

Theorem 6.5. *Given integers* $e \geq 0$ *and* $m \geq e$*, consider the contact locus*

$$\mathrm{Cont}^e(K_{X'/X})_m = \{\gamma' \in X'_m \mid \mathrm{ord}_{\gamma'}(K_{X'/X}) = e\}.$$

If $m \geq 2e$*, then* $\mathrm{Cont}^e(K_{X'/X})_m$ *is a union of fibres of* $\mu_m\colon X'_m \longrightarrow X_m$*, each of which is isomorphic to an affine space* $\mathbb{A}^e$*. Moreover, if*

$$\gamma', \gamma'' \in \mathrm{Cont}^e(K_{X'/X})_m$$

lie in the same fibre of μ_m*, then they have the same image in* X'_{m-e}.

As an application of Theorems 6.1 and 6.2, one gives in [18] a simple proof of the following result of Mustaţă [37] describing the log canonical threshold in terms of the geometry of the space of jets.

Theorem 6.6. *Let* X *be a smooth complex variety and* Y *be a closed subscheme of* X *defined by the nonzero ideal sheaf* I_Y*. Let* X_m *and* Y_m *be the spaces of* m^{th} *order jets of* X *and* Y*, respectively. If* $c = \mathrm{lc}(X, Y)$*, then*

(i) *For every m we have* $\operatorname{codim}(Y_m, X_m) \geq c \cdot (m+1)$. *More generally, if $W \subset X_\infty$ is an irreducible closed cylinder that does not dominate X, then* $\operatorname{codim}(W, X_\infty) \geq c \cdot \operatorname{val}_W(I_Y)$.

(ii) *If m is sufficiently divisible, then* $\operatorname{codim}(Y_m, X_m) = c \cdot (m+1)$.

(iii) *We have* $c = \lim_{m\to\infty} \frac{\operatorname{codim}(Y_m, X_m)}{m+1}$.

The above results relating divisorial valuations with the space of arcs can be used to study more subtle invariants of singularities of pairs. Let Y be a closed subscheme of the smooth variety X, and let λ be a positive real number. We associate a numerical invariant to the pair $(X, \lambda \cdot Y)$ and to an arbitrary nonempty closed subset $B \subseteq X$, as follows.

Consider a divisorial valuation of the form val_E with center $c_X(E)$ in X (the center is the image of E in X). The *log discrepancy* of the pair $(X, \lambda \cdot Y)$ along E is

$$a(E, X, \lambda \cdot Y) = k_E + 1 - \lambda \cdot \operatorname{val}_E(I_Y),$$

where I_Y is the ideal of Y in X. Note that if $Y = \emptyset$, we recover the log discrepancy of val_E. The idea is to measure the singularities of the pair $(X, \lambda \cdot Y)$ using the log discrepancies along divisors with center contained in B.

Definition 6.7. Let $B \subset X$ be a nonempty closed subset. The minimal log discrepancy of $(X, \lambda \cdot Y)$ over B is defined by

$$\operatorname{mld}(B; X, \lambda \cdot Y) := \inf_{c_X(E) \subseteq B} \{a(E; X, \lambda \cdot Y)\}. \tag{6}$$

Remark 6.8. One can show that $\operatorname{mld}(B; X, \lambda \cdot Y)$ is either $-\infty$ or a nonnegative real number. In fact, $\operatorname{mld}(B; X, \lambda \cdot Y) \neq -\infty$ if and only if there is an open neighborhood U of B such that $\operatorname{lc}(U, U \cap Y) \geq \lambda$. An important fact about minimal log discrepancies is that they can be computed using a log resolution of $(X, B \cup Y)$, see [1].

The following theorem of Ein, Mustaţă and Yasuda [22] gives a description of minimal log-discrepancies in term of the geometry of the space of arcs.

Theorem 6.9. *Let B be a nonempty, proper closed subset of X, and let $\pi : X_\infty \longrightarrow X$ be the projection map. For every proper closed subscheme Y of X and for every λ and $\tau \in \mathbb{R}_+$ we have* $\operatorname{mld}(B; X, \lambda \cdot Y) \geq \tau$ *if and only if for every irreducible closed cylinder $W \subseteq \pi^{-1}(B)$ we have*

$$\operatorname{codim}(W, X_\infty) \geq \lambda \cdot \operatorname{val}_W(I_Y) + \tau.$$

The above theorem can be applied to study the behavior of singularities of pairs under restriction to a divisor. This is useful whenever one wants to do induction on dimension. Suppose that D is a smooth divisor on X. We want to relate the singularities of $(X, \lambda \cdot Y)$ with those of $(D, \lambda \cdot Y|_D)$. The adjunction formula suggests that the precise relation should be between $(X, D + \lambda \cdot Y)$ and $(D, \lambda \cdot Y|_D)$. The precise formula is the content of the following theorem from [22].

Theorem 6.10. *Let D be a smooth divisor on the smooth variety X and let B be a nonempty proper closed subset of D. If Y is a closed subscheme of X such that $D \not\subseteq Y$, and if $\lambda \in \mathbb{R}_+$, then*

$$\mathrm{mld}(B; X, D + \lambda \cdot Y) = \mathrm{mld}(B; D, \lambda \cdot Y|_D).$$

Remark 6.11. The notion of minimal log discrepancy plays an important role in the Minimal Model Program. It can be defined under weak assumptions on the singularities of X: one requires only that X is normal and $\mathbb{Q}$-Gorenstein. Kollár and Shokurov have conjectured the statement of Theorem 6.10 with the assumption that X and D are only normal and $\mathbb{Q}$-Gorenstein. It is easy to see that the inequality "$\leq$" holds in general, and the opposite inequality is known as Inversion of Adjunction (see [32] and [33] for a discussion of this conjecture and related topics). Theorem 6.10 has been generalized in [21] to the case when both X and D are normal locally complete intersections.

The interpretation of minimal log discrepancies in terms of spaces of arcs gives also the following semicontinuity statement. This was conjectured for an arbitrary (normal and $\mathbb{Q}$-Gorenstein) variety X by Ambro and Shokurov, see [1]. The statement below, due to Ein, Mustaţă and Yasuda [22] has been generalized to the case of a normal locally complete intersection variety in [21].

Theorem 6.12. *If X is a smooth variety and if Y is a closed subscheme of X, then for every $\lambda \in \mathbb{R}_+$, the function on X defined by $x \to \mathrm{mld}(x; X, \lambda \cdot Y)$ is lower semicontinuous.*

We end with a result that translates properties of the minimal log discrepancy over the singular locus of a locally complete intersection variety into geometric properties of its spaces of jets.

Theorem 6.13. *Let X be a normal locally complete intersection variety of dimension n.*

(i) *X_m has pure dimension $n(m+1)$ for every m (and in this case X_m is also a locally complete intersection) if and only if* $\mathrm{mld}(X_{\mathrm{sing}}; X, \emptyset) \geq 0$ (*this says that X has* log canonical singularities).

(ii) *X_m is irreducible for every m (and in this case it is also reduced) if and only if* $\mathrm{mld}(X_{\mathrm{sing}}; X, \emptyset) \geq 1$ (*this says that X canonical singularities*).

(iii) *X_m is normal for every m if and only if* $\mathrm{mld}(X_{\mathrm{sing}}; X, \emptyset) > 1$ (*this says that X has* terminal singularities).

(iv) *In general, we have* $\mathrm{codim}((X_m)_{\mathrm{sing}}, X_m) \geq \mathrm{mld}(X_{\mathrm{sing}}; X, \emptyset)$ *for every m.*

Remark 6.14. The description in (ii) above was first proved in [38]. Note that since X is in particular Gorenstein, it is known that X has canonical singularities if and only

if it has rational singularities. All the statements in the above theorem were obtained in [22] and [21] combining the description of minimal log discrepancies in terms of spaces of arcs and Inversion of Adjunction.

7. Invariants in positive characteristic

Several invariants have been recently introduced in positive characteristic using the Frobenius morphism, invariants whose behavior is formally very similar to the ones we have discussed in characteristic zero. Moreover, there are interesting results and conjectures involving the comparison between the two sets of invariants via reduction mod p.

As in the case of singularities of pairs (X, Y) in characteristic zero, one can develop the theory under very mild assumptions on the ambient variety X (in fact, the positive characteristic theory does not even need the assumption that X is $\mathbb{Q}$-Gorenstein). For this one needs to use the full power of the theory of tight closure of Hochster and Huneke [26]. However, the definitions become particularly transparent if we assume X nonsingular. Therefore, in accord with the setup in the previous sections, we will make this assumption. The theory we present here is due to Hara and Yoshida [25] building on previous work of Hara, Smith, Takagi and Watanabe.

We work in the local setting with a regular local ring $(R, \mathfrak{m}, k)$ of characteristic $p > 0$. Let $n = \dim(R)$ and let E be the top local cohomology module of R, $E = H^n_{\mathfrak{m}}(R)$. If $x_1, \ldots, x_n$ generate $\mathfrak{m}$, then

$$E \simeq R_{x_1 \ldots x_n} \Big/ \sum_{i=1}^{n} R_{x_1 \ldots \widehat{x_i} \ldots x_n}. \tag{7}$$

The Frobenius morphism on R induces a Frobenius morphism F_E on E that via the isomorphism (7) takes the class of $u/(x_1 \ldots x_n)^d$ to the class of $u^p/(x_1 \ldots x_n)^{pd}$.

We want to study the singularities of the pair (X, Y), where $X = \operatorname{Spec}(R)$ and Y is defined by a nonzero ideal $\mathfrak{a}$. For every $r \geq 0$ and every $e \geq 1$, we put

$$Z_{r,e} := \ker(\mathfrak{a}^r F_E^e) = \{w \in E \mid h F_E^e(w) = 0 \text{ for all } h \in \mathfrak{a}^r\}.$$

Given a nonnegative real number λ, the *test ideal* of the pair $(X, \lambda \cdot Y)$ is

$$\tau(X, \lambda \cdot Y) := \operatorname{Ann}_R\Big(\bigcap_{e \geq 1} Z_{\lceil \lambda p^e \rceil, e}\Big).$$

Here $\lceil \alpha \rceil$ denotes the smallest integer that is $\geq \alpha$.

As Hara and Yoshida show in [25], the test ideals $\tau(X, \lambda \cdot Y)$ enjoy formal properties similar to those of the multiplier ideals $\mathcal{J}(X, \lambda \cdot Y)$ in characteristic zero. In particular, we can consider the jumping numbers for the test ideals: these are the λ such that $\tau(X, \lambda \cdot Y) \subsetneq \tau(X, (\lambda - \varepsilon) \cdot Y)$ for every positive ε.

The set of jumping numbers for the test ideals are also eventually periodic with period one. However, two basic properties that for multiplier ideals follow simply from the description in terms of a log resolution are not known for test ideals: it is not known whether every jumping number for the test ideals is rational, and whether in every bounded interval there are only finitely many such jumping numbers. We want to stress that the problem does *not* come from the fact that we do not know, in general, whether such resolutions exist. Even when we have such resolutions, the invariants in characteristic p do not depend simply on the numerical data of the resolution (see Example 7.4 below for the case of the cusp).

There is a more direct description of the set of jumping numbers given by Mustaţă, Takagi and Watanabe in [39]. Suppose that J is a proper ideal of R containing $\mathfrak{a}$ in its radical. For every $e \geq 1$, define $\nu^J(p^e)$ to be the largest r such that $\mathfrak{a}^r$ is not contained in the e^{th} Frobenius power of J

$$J^{[p^e]} := (u^{p^e} \mid u \in J).$$

It is easy to see that $\nu^J(p^e)/p^e \leq \nu^J(p^{e+1})/p^{e+1}$, and the *F-threshold* of $\mathfrak{a}$ with respect to J is defined by

$$c^J(\mathfrak{a}) := \sup_e \frac{\nu^J(p^e)}{p^e}.$$

It is shown in [39] that the set of F-thresholds of $\mathfrak{a}$ (with respect to various J) is precisely the set of jumping numbers for the test ideals of (X, Y). Note that the smallest F-threshold is obtained for $J = \mathfrak{m}$: this is an analogue of the log canonical threshold that was introduced and studied by Takagi and Watanabe in [46].

There are several interesting results and questions relating the invariants in characteristic zero and those obtained via reduction mod p. To keep the notation simple we will work in the following setup. Suppose that $\mathfrak{a}$ is an ideal in $A[x_1, \ldots, x_n]$, where A is the localization of $\mathbb{Z}$ at some integer. Let Y be the subscheme of $X = \mathbb{A}^n_A$ defined by $\mathfrak{a}$. If p is a prime that is large enough, then by reducing mod p and localizing at $(x_1, \ldots, x_n)$ we get a closed subscheme Y_p in $X_p = \operatorname{Spec} \mathbb{F}_p[x_1, \ldots, x_n]_{(x_1,\ldots,x_n)}$ defined by the ideal $\mathfrak{a}_p$.

The multiplier ideals of the pair (X, Y) (more precisely, of its extension to $\mathbb{C}$) can be computed by a log resolution defined over $\mathbb{Q}$. After suitably localizing A we may assume that the multiplier ideals are defined over A, too. The following results relate the reduction mod p of the multiplier ideals with the test ideals. They are due to Hara and Yoshida [25], based on previous work of Hara, Smith, Takagi and Watanabe.

Theorem 7.1. *With the above notation, if $p \gg 0$, then for every λ we have*

$$\tau(X_p, \lambda \cdot Y_p) \subseteq \mathcal{J}(X, \lambda \cdot Y)_p.$$

Note that since our primes are large enough, the log resolution over $\mathbb{Q}$ induces by reduction mod p log resolutions for (X_p, Y_p). The proof of Theorem 7.1 is based on the use of local duality for the reduction mod p of the log resolution. The proof of the

next result is more involved, using the approach of Deligne and Illusie to the positive characteristic proof of the Kodaira Vanishing Theorem.

Theorem 7.2. *With the above notation, for every λ and for every $p \gg 0$ (depending on λ) we have*

$$\tau(X_p, \lambda \cdot Y_p) = \mathcal{J}(X, \lambda \cdot Y)_p.$$

Remark 7.3. We reinterpret the above statements in terms of jumping numbers. For simplicity, we restrict ourselves to the smallest such number: given $\mathfrak{a}$ as above and $p \gg 0$, we want to compare the log canonical threshold c of the pair (X, Y) in some small neighborhood of the origin, with the F-pure threshold $c_p = c^{\mathfrak{m}}(\mathfrak{a}_p)$. Theorem 7.1 implies that for all $p \gg 0$ we have $c \geq c_p$, while Theorem 7.2 implies that $\lim_{p\to\infty} c_p = c$.

Example 7.4. Let $\mathfrak{a}$ be generated by $f = x^2 + y^3$, whose log canonical threshold is $\frac{5}{6}$. Let $p > 3$ be a prime. One can show that if $p \equiv 1 \pmod 3$, then the largest r such that f^r does not lie in (x^{p^e}, y^{p^e}) is given by $\nu(p^e) = \frac{5}{6}(p^e - 1)$ for every $e \geq 1$, so that $c_p = \frac{5}{6}$. On the other hand, if $p \equiv 2 \pmod 3$, then $\nu(p) = \frac{5p-7}{6}$, while $\nu(p^e) = \frac{5p^e - p^{e-1} - 6}{6}$ for $e \geq 2$. Therefore in this case $c_p = \frac{5}{6} - \frac{1}{6p}$.

Conjecture 7.5. For every ideal $\mathfrak{a}$ in $A[x_1, \ldots, x_n]$ there are infinitely many primes p for which the F-pure threshold c_p is equal to the log canonical threshold c.

For a discussion of this conjecture we refer to [39]. We end by mentioning a connection between the positive characteristic invariants and the Bernstein–Sato polynomial. Suppose that $f \in A[x_1, \ldots, x_n]$ is as above. We know that the Bernstein–Sato polynomial $b_f(s)$ has rational roots, and in fact, one can show that one can find an equation (3) as in the definition of $b_f(s)$ with P having rational coefficients. Therefore, after suitably localizing A, we may assume that both b_f and P have coefficients in A and that (3) holds over A. It follows that if p is a large enough prime, we get a similar equation over $\mathbb{F}_p$.

Consider now an ideal J in the ring $\mathbb{F}_p[x_1, \ldots, x_n]_{(x_1,\ldots,x_n)}$, such that f_p lies in the radical of J. Let us apply (3) with $s = \nu^J(p^e)$, the largest integer such that f_p^r is not in $J^{[p^e]}$. Since the ideal $J^{[p^e]}$ is a module over the ring $\mathbb{F}_p[x, \partial_x]$, we deduce that $b_f(\nu^J(p^e)) \equiv 0 \pmod p$. Therefore the functions ν^J give roots of $b_f \bmod p$. Sometimes one can use this observation to find actual roots of b_f.

Example 7.6. Let $f = x^2 + y^3$. We have described in Example 7.4 the function $\nu = \nu^J$ when J is the maximal ideal. If $p \equiv 1 \pmod 3$, then $\nu(p^e) = \frac{5}{6}(p^e - 1)$. The above discussion implies that p divides $b_f(-5/6)$. Since there are infinitely many such primes, we deduce that $-\frac{5}{6}$ is a root of b_f. Similarly, if $p \equiv 2 \pmod 3$, then it follows from the formula for $\nu(p)$ that $-\frac{7}{6}$ is a root of b_f, and from the formula for $\nu(p^e)$, with $e \geq 2$ that -1 is a root of b_f. Therefore we have obtained all roots of the Bernstein–Sato polynomial of f by this method.

A similar picture can be seen in other examples, though at the moment there is no general result in this direction. In [6] this approach was used to describe all the roots of the Bernstein–Sato polynomial of a monomial ideal. It would be very interesting to find a more conceptual framework that would explain the connection between the Bernstein polynomial and the invariants in positive characteristic.

Acknowledgements. We would like to express a special thanks to Rob Lazarsfeld. Many results in this paper were joint work with him. Moreover, we have benefited from many inspiring discussions.

References

[1] Ambro, F., On minimal log discrepancies. *Math. Res. Lett.* **6** (1999), 573–580.

[2] Arnold, V. I., Gusein-Zade, S. M., and Varchenko, A. N., *Singularities of differentiable maps* I, II. Monogr. Math. 82, 83, Birkhäuser, Boston, MA, 1985.

[3] Angehrn, U., and Siu, Y.-T., Effective freeness and point separation for adjoint bundles. *Invent. Math.* **122** (1995), 291–308.

[4] Bernstein, J. N., Analytic continuation of generalized functions with respect to a parameter. *Funkcional. Anal. i Priložen.* **6** (1972), 26–40; English transl. *Funct. Anal. Appl.* **6** (1972), 273–285.

[5] Budur, N., and Saito, M., Multiplier ideals, V-filtration, and spectrum. *J. Algebraic Geom.* **14** (2005), 269–282.

[6] Budur, N., Mustaţă, M., and Saito, M., Roots of Bernstein-Sato polynomials for monomial ideals: a positive characteristic approach. *Math. Res. Lett.*, to appear; math. AG/0505472.

[7] Budur, N., Mustaţă, M., and Saito, M., Bernstein-Sato polynomials of arbitrary varieties. *Compositio Math.*, to appear; math. AG/0408408.

[8] Cheltsov, I. A., On a smooth four-dimensional quintic. *Mat. Sb.* **191** (2000), 139–160; English transl. *Sb. Math.* **191** (2000), 1399–1419,

[9] Corti, A., Singularities of linear systems and 3-fold birational geometry. In *Explicit birational geometry of 3-folds*, London Math. Soc. Lecture Note Ser. 281, Cambridge University Press, Cambridge 2000, 259–312.

[10] de Fernex, T., Ein, L., and Mustaţă, M., Multiplicities and log canonical threshold. *J. Algebraic Geom.* **13** (2004), 603–615.

[11] de Fernex, T., Ein, L., and Mustaţă, M., Bounds for log canonical thresholds with applications to birational rigidity. *Math. Res. Lett.* **10** (2003), 219–236.

[12] Demailly, J.-P., L^2 vanishing theorems for positive line bundles and adjunction theory. In *Transcendental methods in Algebraic Geometry* (Cetraro, Italy, July 1994), Lecture Notes in Math. 1646, Springer-Verlag, Berlin 1996, 1–97.

[13] Demailly, J.-P., Ein, L., and Lazarsfeld, R., A subadditivity property for multiplier ideals. *Michigan Math. J.* **48** (2000), 137–156.

[14] Denef, J., and Loeser, F., Germs of arcs on singular varieties and motivic integration. *Invent. Math.* **135** (1999), 201–232.

[15] Ein, L., Multiplier ideals, vanishing theorem and applications. In *Algebraic Geometry, Santa Cruz 1995*, Proc. Sympos. Pure Math. 62, Amer. Math. Soc., Providence, RI, 1997, 203–219.

[16] Ein, L., and Lazarsfeld, R., A geometric effective Nullstellensatz. *Invent. Math.* **137** (1999), 427–448.

[17] Ein, L., and Lazarsfeld, R., Singularities of theta divisors and the birational geometry of irregular varieties. *J. Amer. Math. Soc.* **10** (1997), 243–258.

[18] Ein, L., Lazarsfeld, R., Mustaţă, M., Contact loci in arc spaces. *Compositio Math.* **140** (2004), 1229–1244.

[19] Ein, L., Lazarsfeld, R., and Smith, K., Uniform bounds and symbolic powers on smooth varieties. *Invent. Math.* **144** (2001), 241–252.

[20] Ein, L., Lazarsfeld, R., Smith, K.,and Varolin, D., Jumping coefficients of multiplier ideals. *Duke Math. J.* **123** (2004), 469–506.

[21] Ein, L., M. Mustaţă, M., Inversion of adjunction for local complete intersection variety. *Amer. J. Math.* **126** (2004), 1355–1365.

[22] Ein, L., Mustaţă, M., and Yasuda, T., Jet schemes, log discrepancies and inversion of adjunction. *Invent. Math.* **153** (2003), 519–535.

[23] Hacon, C., and McKernan, J., On the existence of flips. math. AG/0507597.

[24] Hacon, C., and McKernan, J., Boundedness of pluricanonical maps of varieties of general type. math. AG/0504327.

[25] Hara, N., and Yoshida, K.-i., A generalization of tight closure and multiplier ideals. *Trans. Amer. Math. Soc.* **355** (2003), 3143–3174.

[26] Hochster, M., and Huneke, C., Tight closure, invariant theory and the Briançon-Skoda theorem. *J. Amer. Math. Soc.* **3** (1990), 31–116.

[27] Howald, J., Multiplier ideals of monomial ideals. *Trans. Amer. Math. Soc.* **353** (2001), 2665–2671.

[28] Iskovskikh, V. A., and Manin, Yu. I., Three-dimensional quartics and counterexamples to the Lüroth problem. *Mat. Sb.* **86** (1971), 140–166; English transl. *Math. Sb.* **15** (1972), 141–166.

[29] Kashiwara, M., Vanishing cycle sheaves and holonomic systems of differential equations. In *Algebraic Geometry* (Tokyo/Kyoto, 1982), Lecture Notes in Math. 1016, Springer-Verlag, Berlin 1983, 134–142.

[30] Kashiwara, M., B-functions and holonomic systems. *Invent. Math.* **38** (1976/77), 33–53.

[31] Kollár, J., *Shafarevich maps and automorphic forms*. Princeton University Press, Princeton, NJ, 1995.

[32] Kollár, J., Singularities of pairs. In *Algebraic Geometry, Santa Cruz 1995*, Proc. Sympos. Pure Math. 62, Amer. Math. Soc., Providence, RI, 1997, 221–286.

[33] Kollár, J., (with 14 coauthors), Flips and abundance for algebraic threefolds. *Astérisque* **211** (1992).

[34] Lazarsfeld, R., Positivity in algebraic geometry, I, II. Ergeb. Math. Grenzgeb. (3) 48, 49, Springer-Verlag, Berlin 2004.

[35] Lichtin, B., Poles of $|f(z, w)|^{2s}$ and roots of the B-function. *Ark. Math.* **27** (1989), 283–304.

[36] Malgrange, B., Polynôme de Bernstein-Sato et cohomologie évanescente. In *Analysis and topology on singular spaces, II, III* (Luminy, 1981), *Astérisque* **101–102** (1983), 243–267.

[37] Mustaţă, M., Singularities of pairs via jet schemes. *J. Amer. Math. Soc.* **15** (2002), 599–615.

[38] Mustaţă, M., Jet schemes of locally complete intersection canonical singularities (with an appendix by D. Eisenbud and E. Frenkel). *Invent. Math.* **145** (2001), 397–424.

[39] Mustaţă, M., Takagi, S., and Watanabe, K.-i., F-thresholds and Bernstein-Sato polynomials. In *European Congress of Mathematics* (ECM), Stockolm, Sweden, June 27–July 2, 2004, EMS Publishing House, Zürich 2005.

[40] Pukhlikov, A. V., Birational automorphisms of a four-dimensional quintic. *Invent. Math.* **87** (1987), 303–329.

[41] Pukhlikov, A. V., Birational automorphisms of Fano hypersurfaces. *Invent. Math.* **134** (1998), 401–426.

[42] Pukhlikov, A. V., Birationally rigid Fano hypersurfaces. *Izv. Math.* **66** (2002), 1243–1269.

[43] Saito, M., Sturmfels, B., and Takayama, N., *Gröbner deformations of hypergeometric differential equations*. Algorithms Comput. Math. 6, Springer-Verlag, Berlin 2000.

[44] Siu, Y.-T., Invariance of plurigenera and torsion-freeness of direct image sheaves of pluricanonical bundles. In *Finite or infinite dimensional complex analysis and applications*, Adv. Complex Anal. Appl. 2, Kluwer Acadademic Publ., Dordrecht 2004, 45–83.

[45] Siu, Y.-T., Invariance of plurigenera. *Invent. Math.* **134** (1998), 661–673.

[46] Takagi, S., and Watanabe, K.-i., On F-pure thresholds. *J. Algebra* **282** (2004), 278–297.

[47] Takayama, S., Pluricanonical systems on algebraic varieties of general type. Preprint, 2005.

[48] Varchenko, A., Asymptotic Hodge structures in the vanishing cohomology. *Math. USSR Izv.* **18** (1982), 469–512.

[49] Yano, T., b-functions and exponents of hypersurface isolated singularities. *Proc. Sympos. Pure Math* **40** (1983), 641–652.

Department of Mathematics, University of Illinois at Chicago, 851 South Morgan Street (M/C 249), Chicago, IL 60607-7045, U.S.A.
and
Department of Mathematics, University of California at Irvine, Irvine, CA 92697-3875, U.S.A.
E-mail: ein@math.uic.edu

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
E-mail: mmustata@umich.edu

Rational curves and rational points

Tom Graber*

Abstract. We survey some recent results proving the existence of rational points over one dimensional function fields and finite fields on varieties containing many rational curves. We also consider conjectural extensions of these theorems to higher dimensional function fields.

Mathematics Subject Classification (2000). 14G05.

Keywords. Rationally connected varieties, rational points.

1. Introduction

In the past five years there has been substantial progress giving geometric interpretations and generalizations of classical theorems about C_1 fields. In this lecture I would like to focus on some results about the existence of rational points on certain classes of varieties over function fields of curves and over finite fields. The geometry of rational curves play a central role in this work.

Recent results of de Jong and Starr give hope that in the future we will see a similar development for function fields of higher dimensional varieties, although so far we do not have even a precise conjectural formulation. The program that they and others have proposed raises many interesting questions, most of which are still relatively unexplored.

It is a pleasure to thank my collaborators from whom I learned most of what I know about this material: Joe Harris, Barry Mazur, and Jason Starr, to whom I am particularly grateful for helpful comments on this lecture.

2. Classical results

All the results and conjectures that I want to consider have their origins in two theorems proven in the 1930s which guarantee that over certain fields, hypersurfaces of low degree in projective space possess rational points. To make this precise, we make a definition.

*The author is grateful to the Institut des Hautes Études Scientifiques for hospitality during the writing of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Definition 2.1. A field K is said to be C_1 if every homogeneous polynomial $f \in K[x_0, \ldots, x_n]$ has a nontrivial zero provided that $\deg(f) \leq n$.

The first two known classes of C_1 fields were provided by the following two well-known theorems.

Theorem 2.2 (Tsen). *If k is an algebraically closed field and C is an irreducible curve over k with function field K, then K is a C_1 field.*

Theorem 2.3 (Chevalley). *Finite fields are C_1.*

There are two natural ways to search for generalizations of these results: we can try to extend the class of fields considered, or extend the class of varieties. There are substantial generalizations in both directions. We will start by considering the second question, namely, for which types of varieties can we guarantee that they will possess points over finite fields or over one dimensional function fields? At the expense of making less than optimal statements, we will always restrict our attention to smooth projective varieties. In this context, there are many natural generalizations of the class of hypersurfaces above. The two most obvious ones for an algebraic geometer before 1992 were the following.

Remark 2.4. If X is a smooth hypersurface in $\mathbb{P}^n$ of degree d, then the following conditions are equivalent.

- $d \leq n$.
- X is Fano, that is, the canonical line bundle K_X has ample dual.
- $h^i(X, \mathcal{O}_X) = 0$ for all $i > 0$.

These naturally gives rise to four questions, do Tsen's or Chevalley's theorems generalize to either of these classes of varieties? The answer to one of these questions has been known for some time.

Theorem 2.5 (Katz). *If X is a projective variety over $\mathbb{F}_q$ such that $h^i(X, \mathcal{O}_X) = 0$ for all $i > 0$, then X has an $\mathbb{F}_q$-point.*

The other three have all been (nearly) settled in the new millennium, in large part due to a fundamental shift in viewpoint coming from a third geometric condition – *rational connectivity* – which is also equivalent for smooth hypersurfaces to the inequality $d \leq n$.

3. Rationally connected varieties

The basic theory of rational connectivity was developed independently by Campana [C] and Kollar, Miyaoka, and Mori [KMM]. We will give a quick review of some of the main points. Throughout, we work over a ground field k. To simplify our

discussion we will restrict our attention to smooth projective varieties, and we will take k to be an uncountable algebraically closed field. A careful reader will notice that this last hypothesis will not be satisfied in many of the applications we refer to later. To avoid this, one must either make more careful definitions, or first base change to such a field. For a good treatment of the foundations of the theory of rationally connected varieties, we refer to [K] or [D].

Given a projective variety X, by a *rational curve* in X, we mean the image of a morphism $f\colon \mathbb{P}^1 \to X$.

Definition 3.1. A smooth projective variety X is *rationally connected* (RC) if two general points of X can be connected by a rational curve.

If $k = \mathbb{C}$ this notion is equivalent to many other conditions guaranteeing the existence of an abundance of rational curves in X. Since these other conditions can fail to be equivalent to rational connectedness over other fields (or for singular varieties), they have their own names.

Definition 3.2. X is *rationally chain connected* (RCC) if any two points of X can be joined by a chain of rational curves.

Definition 3.3. X is *separably rationally connected* (SRC) if there exists a morphism $f\colon \mathbb{P}^1 \to X$ such that $f^*(TX)$ is an ample vector bundle.

While this last definition appears quite different from the previous two, it is straightforward to verify that such a morphism deforms so freely that it guarantees the existence of a rational curve through two general points. Combining this with the elementary fact that every irreducible component of a degeneration of a rational curve is rational, we obtain the implications

$$\text{SRC} \implies \text{RC} \implies \text{RCC}.$$

In characteristic zero, by more delicate arguments, one can reverse these implications. The equivalence of these notions is quite convenient, because it is easy to see that the property of being rationally chain connected is closed in families, while that of being separably rationally connected is open. As a result, we find that in characteristic zero, rational connectivity is a deformation invariant property. This is one reason why the notion is particularly well suited to the classification theory of higher dimensional varieties.

The following beautiful theorem, which even over the complex numbers can be proven at this time only by making use of results in characteristic p, was established in [C] and [KMM].

Theorem 3.4. *Fano varieties are rationally chain connected.*

As a result, one can try to address questions about Fano varieties by studying the geometry of rational curves. Indeed, rational connectivity seems to be a more useful notion for questions about the existence of rational points.

4. Rational points on rationally connected varieties

For both finite fields and function fields, the study of rational points often takes as its starting point a more geometric interpretation of the notion of rational point. The very different methods used in the two cases reflect the difference in this geometric interpretation. We will start with the case of function fields.

If we let K be the function field of a smooth curve C defined over an algebraically closed field k, then given a projective variety X_K over K, we can always find a C model for X_K. That is, we can find a projective variety X over k together with a morphism $\pi : X \to C$ such that the base change to Spec K is X_K. A rational point of X_K then corresponds exactly to a section of π. Thus the geometry of rational points over K is intimately connected to the geometry of curves in higher dimensional varieties. From this point of view, then, the notion of rational connectivity is particularly appealing, since it is also directly connected with the same type of geometric objects. Moreover, it is not difficult to see that the variety X_K is geometrically rationally connected if and only if a general fiber of the morphism π is rationally connected.

Thus, the following theorem, which is proven in characteristic zero in [GHS] and in positive characteristic in [dJS] is a natural generalization of Tsen's theorem.

Theorem 4.1. *If C is a smooth curve over an algebraically closed field, and $\pi : X \to C$ is a proper morphism whose general fiber is separably rationally connected, then π admits a section.*

Combining with Theorem 3.4 we get a partial answer to one of the questions raised in Section 2.

Corollary 4.2. *If K is the function field of a curve defined over an algebraically closed field of characteristic zero, then every Fano variety over K has a rational point.*

To my knowledge the analogous question in positive characteristic remains open, since Fano varieties are not known to be separably rationally connected.

The proof of Theorem 4.1 is based on the geometry of Kontsevich's space of stable maps, $\overline{M}_g(X)$. This space, which compactifies the space of smooth curves in X by allowing them to degenerate to morphisms from nodal curves, was introduced into mathematics in order to study ideas coming from string theory. It is extremely useful in this context, however, because it has two advantages over the traditional compactifications of the space of curves in X. Its deformation theory is extremely simple and it has an obvious functorial property – given a morphism of varieties $\pi : X \to Y$, we get an induced morphism $\overline{M}(\pi) : \overline{M}_g(X) \to \overline{M}_g(Y)$. In the setting of Theorem 4.1, the point is that via degeneration methods, one can produce a curve B in X such that B corresponds to a smooth point of $\overline{M}_g(X)$ at which the differential of $\overline{M}(\pi)$ is surjective. As this last statement involves only the tangent spaces to the moduli spaces, it can be verified using elementary deformation theory, but nonetheless

it is very powerful, since it guarantees that $\overline{M}(\pi)$ dominates at least one irreducible component of $\overline{M}_g(C)$. Thus, the proof is completed by finding a suitable degeneration of a morphism of curves, which can be done by elementary techniques.

In [GHMS], by making a careful study of the maps $\overline{M}(\pi)$ for general morphisms of varieties π it is shown that in a certain sense the property of rational connectivity is universal for the problem of finding rational points over function fields of curves. In other words, we cannot really hope to find a larger class of varieties which will always possess a point over function fields than the class given by Theorem 4.1.

In particular, there exist varieties X over $K(C)$ such that $h^i(X, \mathcal{O}_X) = 0$ for all $i > 0$ but without a rational point. We remark that while the methods of [GHMS] prove the existence of such a variety, they are not effective and give little insight into how to actually write one down. In [Laf] G. Lafon produces an explicit Enriques surface over $\mathbb{Q}(t)$ with no rational point even over $\mathbb{C}((t))$.

Turning our attention to finite fields, the theory has a rather different flavor. The geometric interpretation of an $\mathbb{F}_q$-point of a variety X is that it is a fixed point of the action of the Frobenius morphism. This point of view is very useful in connection with cohomological conditions, since one can use Lefschetz type theorems that relate fixed points to cohomological data. In particular, Theorem 2.5 was proven by establishing the analogue for the Frobenius morphism of the holomorphic Lefschetz fixed point theorem – that the number of points of a smooth projective variety over $\mathbb{F}_q$ is congruent modulo p to the alternating sum of the traces of Frobenius on the cohomology groups $H^i(X, \mathcal{O}_X)$. Since the trace of Frobenius on H^0 is obviously 1 the existence result follows immediately, as well as the sharper statement that the number of $\mathbb{F}_q$-points on X is congruent to 1 modulo p.

In [E], H. Esnault shows how it is possible to relate rational curves to this circle of ideas. She observes that an immediate consequence of rational chain connectivity is that the Chow group of zero cycles is $\mathbb{Z}$. A method introduced by Bloch (cf. [B]) can then be used to control the eigenvalues of the action of Frobenius on the étale (or crystalline) cohomology of X. This is done by thinking of the diagonal $\Delta \subset X \times X$ as a zero cycle in $X_{K(X)}$ and using the triviality of the Chow group to move it via rational equivalence to $[\{p\} \times X] \cup \Gamma$ where Γ is a cycle whose projection to the second factor is not dominant. Interpreting this as the decomposition of the identity map on the cohomology of X into the projection onto H^0 plus the projection onto the rest, together with formal (but deep) properties of étale cohomology then yield that the eigenvalues of the action of Frobenius on $H^i_{et}(X, \mathbb{Q}_l)$ are divisible by q for all $i > 0$. Now using the Lefschetz–Verdier trace formula, she concludes the following theorem.

Theorem 4.3. *If X is a smooth projective variety over $\mathbb{F}_q$ with $Ch_0(X \times \overline{K(X)}) = \mathbb{Z}$, then X has a rational point.*

This has as immediate corollaries:

Corollary 4.4. *Every smooth, projective, rationally chain connected variety over a finite field has a rational point.*

Corollary 4.5. *Fano varieties over finite fields have rational points.*

As before, one actually gets a sharper statement – the number of rational points is congruent to 1 mod q.

5. Higher rational connectivity

We would now like to consider some more speculative material about the existence of points on varieties over more complicated fields. We will describe (in very rough terms) a program of de Jong and Starr, and following them, we will consider here only fields of the form $K(X)$ for X a complex variety. We encourage the reader to consider analogous questions over other classes of fields, where it is easy to make similar speculations.

Long before any of the work described here, Serge Lang gave an entirely different generalization of Tsen's theorem. Following him, we first generalize the definition of C_1.

Definition 5.1. A field K is called C_r if every homogeneous polynomial f in $K[x_0, \ldots, x_n]$ has a nontrivial zero provided $d^r \leq n$.

Then Lang proves the following theorem.

Theorem 5.2. *If X is a variety of dimension r over an algebraically closed field, then the function field $K(X)$ is C_r.*

It is natural to ask whether there is a common generalization of Theorem 4.1 and Theorem 5.2. In other words, can we single out a class of abstract varieties which generalizes the class of degree d hypersurfaces in $\mathbb{P}^n$ with $d^r \leq n$ such that they will always have a rational point over function fields of dimension r? So far there are few positive results, but there is an analogy which is extremely tantalizing and which takes as its starting point the observation that rational connectivity is just the usual topological notion of path connectedness with the interval replaced by $\mathbb{P}^1$. If we systematically translate between topology and algebraic geometry using fiber bundle as the analogue of family and the interval as the analogue of $\mathbb{P}^1$, then Theorem 4.1 translates into the elementary topological fact that over a one dimensional manifold, any fiber bundle with connected fiber admits a section.

This topological fact has a straightforward generalization to higher dimensional bases. Namely, if $\phi\colon X \to M$ is a fibration with M an r-dimensional manifold and fiber F such that $\pi_i(F) = 0$ for all $i < r$, then ϕ admits a section. This suggests the hope that we could try to define some notion of higher rational connectedness and prove a theorem that r-rationally connected varieties over r dimensional function fields have rational points.

Unfortunately, such a theorem cannot hold. Certainly any generalization of the class of low degree hypersurfaces will include projective space itself (the case $d = 1$.)

It is well known, however, that over higher dimensional bases, there are families all of whose fibers are isomorphic to $\mathbb{P}^n$ but which do not admit a section. The obstruction to the existence of a section lies in the Brauer group of the base of the family. While this might dash one's hope, de Jong and Starr have taken the optimistic view that at least in dimension 2, the Brauer group might be the only obstruction.

In particular, the Brauer class associated to a family of projective spaces admits a generalization to this situation. Given a morphism $\phi\colon X \to B$, one gets a sequence

$$\mathrm{Pic}(X_K) \to \mathrm{Pic}(X_{\overline{K}})^G \to \mathrm{Br}(K).$$

If X has a K-point, the rightmost map in this sequence vanishes, thus the nonvanishing of this map is an obstruction to the existence of a K-point. We will refer to this as the Brauer obstruction. We remark that this obstruction will always vanish for the projection to B of a hypersurface of dimension at least 3 in $\mathbb{P}^n \times B$. Now we can state the

Metaconjecture. There exists a notion of rationally simply connected such that for any morphism $\pi\colon X \to B$ with B a complex surface and with rationally simply connected general fiber, ϕ admits a rational section if and only if the Brauer obstruction vanishes. Moreover, for smooth hypersurfaces, this notion should agree with the condition that $d^2 \leq n$.

There is a natural guess for how to go about formulating what it means for a variety X to be rationally simply connected. For a topological space, simple connectivity means that the space of loops is path connected, or equivalently that the space of paths between any two fixed points is path connected. By again replacing paths with rational curves, we would arrive at a provisional definition that the space of rational curves joining two general points of X should be rationally connected. Unfortunately, this is impossible, since the space of rational curves joining two points will not even be connected – a rational curve has a discrete invariant, its homology class, which is invariant under deformation. The most we can ask for is for the space of rational curves *of fixed topological type* joining two general points to be rationally connected. Finally, it seems more reasonable to ask for this only in sufficiently positive homology classes since rational connectedness itself is really only a condition on the high degree curves. There are various possible meanings of "sufficiently positive" and the optimal choice is not clear. At least if $\mathrm{Pic}(X) = \mathbb{Z}$ (which will be the case for smooth hypersurfaces), the condition is unambiguous.

Unfortunately, this does not seem to be enough, since de Jong and Starr are able to prove that smooth hypersurfaces of degree d in $\mathbb{P}^n$ satisfy this constraint provided that $d^2 \leq n + 1$ which in this case is too good of a theorem, since Lang's result is sharp. To repair this, they propose that there is an additional condition which might be seen as an analogue of the condition appearing in the definition of separable rational connectivity, namely the existence of a curve in the space of curves satisfying certain positivity properties. This condition is satisfied for hypersurfaces of the desired degree range. We refer the reader to their preprints for the precise conditions they use.

They also propose a strategy for establishing that families of such varieties over surfaces possess a rational section provided that the Brauer obstruction vanishes. To date, they are able to carry out this program only under extremely restrictive hypotheses, which in particular rule out almost all families of hypersurfaces of degree greater than 2. Nonetheless, the work they have done lends tremendous credibility to the belief that there should be theorems in this direction, providing a geometric explanation for (at least the $r = 2$ case of) Lang's theorem. Moreover, their existing results apply to families of Grassmannians and give a new proof of de Jong's period-index theorem.

At this point, one might be inclined to believe that we would at least have a reasonably clear idea of how to generalize this type of conjecture to higher rational connectivity, but there are further issues to confront in order to do this. Since rational connectivity is a property depending only on the birational equivalence class of a variety, in defining rationally simply connected, we did not have to choose a particular compactification of the space of rational curves joining two general points. However, under any definition currently considered, the property of being rationally simply connected is not invariant under birational modification (and should not be), so to generalize these ideas to higher dimension, one needs either to fix a choice of compactification or find a different framework in which to discuss these questions. This direction is wide open and seems likely to provide interesting – and difficult – questions for some time to come.

References

[B] Bloch, S., *Lectures on algebraic cycles.* Duke University Mathematics Series IV, Durham, N.C., 1980.

[C] Campana, F., Connexité rationnelle des variétés de Fano. *Ann. Sci. École Norm. Sup.* **25** (1992), 539–545.

[D] Debarre, O., *Higher dimensional algebraic geometry.* Universitext, Springer-Verlag, New York 2001.

[E] Esnault, H., Varieties over a finite field with trivial Chow group of 0-cycles have a rational point. *Invent. Math.* **151** (2003), 187–191.

[GHMS] Graber, T., Harris, J., Mazur, B., and Starr, J., Rational connectivity and sections of families over curves. *Ann. Sci. École Norm. Sup.*, to appear.

[GHS] Graber, T., Harris, J., and Starr, J., Families of rationally connected varieties. *J. Amer. Math. Soc.* **16** (2003), 57–67.

[dJS] de Jong, A. J., and Starr, J., Every rationally connected variety over the function field of a curve has a rational point. *Amer. J. Math.* **125** (2003), 567–580.

[K] Kollár, J., *Rational curves on algebraic varieties*. Ergeb. Math. Grenzgeb. (3) 32, Springer-Verlag, Berlin 1996.

[KMM] Kollár, J., Miyaoka, Y., and Mori, S., Rationally connected varieties. *J. Algebraic Geom.* **1** (1992), 429–448.

[L] Lang, S., On quasi-algebraic closure. *Ann. of Math.* **55** (1952), 373–390.

[Laf] Lafon, G., Une surface d'Enriques sans point sur $\mathbb{C}((t))$. *C. R. Math. Acad. Sci. Paris* **338** (2004), 51–54.

[T] Tsen, C., Zur Stufentheorie der quasialgebraisch-Abgeschlossenheit kommutativer Körper. *J. Chinese Math.* **1** (1936), 81–92.

Department of Mathematics, California Institute of Technology, Pasadena, CA 91125, U.S.A.
E-mail: graber@caltech.edu

Rigidity of rational homogeneous spaces

Jun-Muk Hwang

Abstract. Rigidity questions on rational homogeneous spaces arise naturally as higher dimensional generalizations of Riemann's uniformization theorem in one complex variable. We will give an overview of some results obtained in this area by the study of minimal rational curves and geometric structures defined by the their tangent directions.

Mathematics Subject Classification (2000). Primary 14J45; Secondary 32M10, 32G05.

Keywords. Uniformization, rational homogeneous space, minimal rational curves.

1. Introduction

Riemann's uniformization theorem in one complex variable says that the three basic Riemann surfaces, namely, the Riemann sphere $\mathbb{P}_1$, the complex plane $\mathbb{C}$ and the unit disc Δ, exhaust all simply connected Riemann surfaces. Finding an analog of this result in higher dimensions has been one of the main themes of research in complex geometry. A number of approaches from different view-points have been developed for this. Our approach here is to regard Riemann's uniformization theorem as a characterization of the three basic Riemann surfaces. From this approach, the uniformization problem in several complex variables is to find suitable conditions which characterize some basic classes of complex manifolds generalizing the three basic Riemann surfaces. The most natural higher-dimensional analogs of the three basic Riemann surfaces are Hermitian symmetric spaces. Thus the uniformization problem in several complex variables leads to the study of rigidity of Hermitian symmetric spaces.

A nice survey of results on the rigidity of Hermitian symmetric spaces obtained up to early 1990s was given in Siu's article 'Uniformization in Several Complex Variables' [28]. As one can see in [28], the methods employed are quite different depending on whether the Hermitian symmetric spaces are of compact type, of Euclidean type, or of non-compact type. The methods used for compact type have close connection with algebraic geometry, while those for Euclidean or non-compact types are closer to differential geometry. Since we have little expertise for the cases of Euclidean type or non-compact type, we will restrict our discussion to the case of compact type. From the view-point of algebraic geometry, it is more natural to consider a larger class of complex manifolds, the rational homogeneous spaces, which include Hermitian symmetric spaces of compact type. We will discuss rigidity problems of

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

rational homogeneous spaces, especially those related to the problems considered in Siu's survey. Our aim is to report the progress on these problems made after 1990.

As explained in Siu's survey, there are two approaches to the uniformization in several complex variables, one via topological conditions and the other via curvature conditions. The approach by topological conditions asks under what additional condition a complex manifold diffeomorphic to a rational homogeneous space is biholomorphic to it. The natural additional condition one can consider is the Kähler condition, the Moishezon condition or the deformation condition. Regarding the first two conditions, little new development was made after [28]. On the other hand, there has been much progress under the deformation condition. This development will be discussed in Section 3. The approach by curvature conditions is to characterize rational homogeneous spaces as manifolds with positive curvature in a suitable sense. There are various ways to impose curvature conditions, but all reasonable assumptions contain the positivity of the anti-canonical bundle. Complex manifolds with positive anti-canonical bundles are called Fano manifolds. Rational homogeneous spaces are just homogeneous Fano manifolds. Thus the question is to characterize homogeneous spaces among Fano manifolds by certain curvature properties of the tangent bundle. Typical examples are the Hartshorne conjecture, the Frankel conjecture and the generalized Frankel conjecture whose solutions by Mori [26], Siu-Yau [29], and Mok [23] were surveyed in [28]. In Section 4, we will discuss the Campana–Peternell conjecture, which generalizes these three results in the context of rational homogeneous spaces.

The results we will discuss on the rigidity problem both under the deformation condition and under the curvature condition depend on the study of minimal rational curves of Fano manifolds. This study originated from Mori's solution of the Hartshorne conjecture [26] which concerns the projective space. To handle problems on general rational homogeneous spaces other than the projective space, it is important to study the tangent directions of minimal rational curves, more precisely, the variety of minimal rational tangents (see Section 2 for definition). We will start with a discussion of this concept and related matters, before we discuss specific rigidity problems in Section 3 and Section 4. The methods of the variety of minimal rational tangents in the study of Fano manifolds have many aspects and applications other than those related to rigidity problems of rational homogeneous spaces. Here we will concentrate only on those aspects directly related to rational homogeneous spaces. For the other aspects, we refer the reader to [11], [14] and [18].

Some comments on the notation. For a vector space V, its projectivization $\mathbb{P}V$ is the set of 1-dimensional subspaces of V. For a complex manifold X and $x \in X$, the holomorphic tangent space of X at x will be denoted by $T_x(X)$ and the holomorphic tangent bundle of X will be denoted by $T(X)$.

Acknowledgment. I would like to thank Ngaiming Mok for valuable comments.

2. Geometric structures arising from minimal rational curves

To handle the rigidity questions, we need to show that a given complex manifold with certain additional conditions is a rational homogeneous space. Thus a common problem in these questions is how to recognize rational homogeneous spaces. Since we are interested in algebro-geometric conditions, we should be able to recognize them in terms of algebro-geometric data. Some special cases, like the projective spaces or the hyperquadrics, can be handled by the properties of certain linear systems. However such approaches are hard to generalize to other rational homogeneous spaces.

To get a hint on this problem, we should look at previous results in uniformization problems concerning rational homogeneous spaces more general than projective spaces or hyperquadrics. Very few results of this type are known. In Siu's survey [28], we found only one result of this type, namely, Mok's solution [23] of the generalized Frankel conjecture, which we will discuss again in Section 4. In Mok's work, after a deformation of the metric by the heat equation, one is given a compact complex manifold X with a Kähler metric g of positive Ricci curvature and of nonnegative holomorphic bisectional curvature, and one has to show that X is a symmetric space. For this, Mok constructed a distinguished subvariety $\mathcal{C}$ of the projectivized tangent bundle $\mathbb{P}T(X)$ which is a proper subvariety unless X is a projective space, and showed that $\mathcal{C}$ is invariant under the holonomy action of the Riemannian metric g. This implies that X is symmetric by Berger's theorem on Riemannian holonomy.

How should we formulate an algebro-geometric analog of Mok's argument? Luckily, Mok's construction of the distinguished subvariety $\mathcal{C} \subset \mathbb{P}T(X)$ is essentially algebraic. His construction can be generalized to arbitrary Fano manifolds as follows.

Let X be a Fano manifold. By Mori [26] X is covered by rational curves. Let us fix an irreducible component $\mathcal{K}$ of the space of rational curves whose members sweep out an open subset of X and have minimal degree with respect to the anticanonical bundle. Members of $\mathcal{K}$ are called minimal rational curves. Whenever we mention minimal rational curves below, we tacitly assume that a choice of $\mathcal{K}$ is made. For a general point $x \in X$, let $\mathcal{K}_x$ be the normalization of the subvariety of $\mathcal{K}$ parametrizing members of $\mathcal{K}$ passing through x. In [26], it is proved that $\mathcal{K}_x$ is a projective manifold. [20] showed that each member of $\mathcal{K}_x$ is immersed at x. Thus there exists a morphism

$$\tau_x : \mathcal{K}_x \longrightarrow \mathbb{P}T_x(X),$$

called the tangent morphism, assigning to each curve its tangent direction at x. Its image $\tau_x(\mathcal{K}_x)$ is called the variety of minimal rational tangents at x and denoted by $\mathcal{C}_x$. Let $\mathcal{C}$ be the closure of the union of $\{\mathcal{C}_x,\ \text{general } x \in X\}$ in $\mathbb{P}T(X)$. $\mathcal{C}$ is called the variety of minimal rational tangents. The variety $\mathcal{C}$ constructed in [23] on a Fano manifold X with nonnegative holomorphic bisectional curvature coincides with the variety of minimal rational tangents on X. It is helpful to look at some examples.

Example 1. Let X be the complex projective space $\mathbb{P}_n$. Minimal rational curves on X are just lines on the projective space. For any point $x \in X$ and any tangent direction

$\alpha \in \mathbb{P}T_x(X)$, there exists a line through x in the direction of α. Thus $\mathcal{C}_x = \mathbb{P}T_x(X)$ and $\mathcal{C} = \mathbb{P}T(X)$.

Example 2. Let X be the n-dimensional hyperquadric $\boldsymbol{Q}_n$ in $\mathbb{P}_{n+1}$. Minimal rational curves on X are just lines of $\mathbb{P}_{n+1}$ which lie on $\boldsymbol{Q}_n$. For a given point $x \in X$, the tangent directions to lines through x lying on $\boldsymbol{Q}_n$ defines a hyperquadric $\mathcal{C}_x \cong \boldsymbol{Q}_{n-2}$ in the projective space $\mathbb{P}T_x(X) \cong \mathbb{P}_{n-1}$.

Example 3. Let X be the Grassmannian $\mathbb{G}(u, v)$ of u-dimensional subspaces of a $(u + v)$-dimensional complex vector space V. There are two universal quotient bundles $\mathcal{U}$ of rank u and $\mathcal{V}$ of rank v such that $T(X) \cong \mathcal{U} \otimes \mathcal{V}$. There is a natural embedding of X into $\mathbb{P}\Lambda^u V$ called the Plücker embedding. Minimal rational curves on X are just lines of $\mathbb{P}\Lambda^u V$ lying on X. It is easy to check that for each $x \in X$, the variety of minimal rational tangents at x, $\mathcal{C}_x \subset \mathbb{P}T_x(X)$, is isomorphic to the set of pure tensors in $\mathbb{P}T_x(X) \cong \mathbb{P}(\mathcal{U}_x \otimes \mathcal{V}_x)$. In other words, $\mathcal{C}_x$ is isomorphic to the Segre embedding of $\mathbb{P}_{u-1} \times \mathbb{P}_{v-1}$.

Now we confront the second part of Mok's argument, which is more subtle: how to use the variety of minimal rational tangents to characterize rational homogeneous spaces. What should be the algebro-geometric analog of Riemannian holonomy and Berger's theorem? Since it is hard to answer this directly, let us approach from a more general setting. Riemannian holonomy is one special part of the general theory of geometric structures in differential geometry. In a broad sense, we can say that a geometric structure is given on a manifold X, if some extra conditions are imposed on the tangent bundle $T(X)$ or its associated fiber bundles. The existence of a distinguished subvariety $\mathcal{C}$ of the projectivized tangent bundle $\mathbb{P}T(X)$ of a complex manifold X can be regarded as a geometric structure on X. For example, the hyperquadric in Example 2 has the structure given by the hypersurface $\mathcal{C} \subset \mathbb{P}T(X)$ which defines a nondegenerate quadratic form on $T_x(X)$ up to a scalar. Such a structure is called a conformal structure. The Grassmannian in Example 3 has the variety of minimal rational tangents isomorphic to the Segre embedding. Such a geometric structure has been studied by differential geometers in connection with twistor theory.

Now that we have a geometric structure arising from minimal rational curves on any Fano manifold, how can we use them to characterize rational homogeneous spaces? A natural approach is by the equivalence problem for such geometric structures, which is the basic problem of E. Cartan's approach to differential geometry. The equivalence problem for the geometric structure is usually a local question. One may wonder whether such a local differential geometric study will be useful in dealing with problems formulated in algebraic geometry. More precisely, let us assume that there exists an analytic open subset $U \subset X$ such that the structure $\mathcal{C}|_U$ is equivalent to that of an analytic open subset of a rational homogeneous space S. What algebro-geometric consequence would this give? For example, can we say that X is biholomorphic to S? The answer is plainly no. A counterexample can be given just by setting X to be the blow-up of S along a subvariety. This example is certainly

uninteresting. There are a number of ways to avoid these inessential problems. One simple way is just to restrict our discussion to complex manifolds of second Betti number 1. This may look like an over-simplification, but there is still a lot to say even under this restriction. From now on let us assume that our complex manifolds have $b_2 = 1$. In this case, we have the following result which shows that the approach via local equivalence of geometric structures is promising.

Theorem 2.1. *Let X be a Fano manifold of $b_2 = 1$ and $\mathcal{C} \subset \mathbb{P}T(X)$ be the variety of minimal rational tangents associated to a family of minimal rational curves. Let S be a rational homogeneous space of $b_2 = 1$ different from $\mathbb{P}_n$ and $\mathcal{C}' \subset \mathbb{P}T(S)$ be the variety of minimal rational tangents. Suppose there exist connected analytic open sets $U \subset X$ and $U' \subset S$ with a biholomorphic map $\varphi\colon U \to U'$ such that its differential $d\varphi\colon \mathbb{P}T(U) \to \mathbb{P}T(U')$ sends $\mathcal{C}|_U$ onto $\mathcal{C}'|_{U'}$. Then φ can be extended to a biholomorphic map $X \cong S$.*

This is a special case of [15], where a similar result was proved for a large class of Fano manifolds including rational homogeneous spaces. The main idea of proof in [15] is the extension of the biholomorphic map φ outside U by an analytic continuation 'along the minimal rational curves'. This extension is possible over the whole manifold X because X is rationally connected by minimal rational curves from $b_2 = 1$. One difficulty of the problem is to prove the univalence of the analytic continuation, where the simply connectedness of X is used crucially.

By Theorem 2.1, the problem of recognizing rational homogeneous spaces of $b_2 = 1$ is reduced to the local equivalence question for the geometric structure defined by the variety of minimal rational tangents. So the question is how to show the existence of the local equivalence map φ. One necessary condition is that for each general point $x \in X$, the variety of minimal rational tangents $\mathcal{C}_x \subset \mathbb{P}T_x(X)$ must be isomorphic to the variety of minimal rational tangents $\mathcal{C}'_s \subset \mathbb{P}T_s(S)$ at a base point $s \in S$. In fact, it is expected that this is a sufficient condition:

Conjecture 2.2. Let S be a rational homogeneous space of $b_2 = 1$ and $\mathcal{C}'_s \subset \mathbb{P}T_s(S)$ be the variety of minimal rational tangents at a base point $s \in S$. Let X be a Fano manifold of $b_2 = 1$ and $\mathcal{C}_x \subset \mathbb{P}T_x(X)$ be the variety of minimal rational tangents at a general point $x \in X$ associated to a family of minimal rational curves. Suppose that $\mathcal{C}'_s \subset \mathbb{P}T_s(S)$ and $\mathcal{C}_x \subset \mathbb{P}T_x(X)$ are isomorphic as projective subvarieties. Then X is biholomorphic to S.

This conjecture is known to be true for a number of cases. When S is the projective space, it follows from the following result proved in [4]:

Theorem 2.3. *Let X be a Fano manifold whose variety of minimal rational tangents at a general point is the whole $\mathbb{P}T_x(X)$. Then X is the projective space.*

The proof in [4] is by a careful study of the singularity of minimal rational curves. When S is the hyperquadric of dimension ≥ 3, Conjecture 2.2 follows from Miyaoka's result [22]. In fact, Miyaoka proved the following stronger result.

Theorem 2.4. *Let X be a Fano manifold of $b_2 = 1$ whose variety of minimal rational tangents at a general point x is a hypersurface in $\mathbb{P}T_x(X)$. Then X is the hyperquadric.*

One special feature of Miyaoka's proof is that it requires a study of rational curves which are not minimal. It would be very interesting if one can prove Theorem 2.4 using only minimal rational curves. Conjecture 2.2 for hyperquadrics can be proved using only minimal rational curves as we will see shortly.

Methods used in [4] or [22] for projective spaces and hyperquadrics are rather special, unrelated to ideas of E. Cartan's equivalence problems. These methods cannot be applied to other rational homogeneous spaces to solve Conjecture 2.2. For general rational homogeneous spaces, the most promising approach seems to be via the techniques used in the equivalence problem for the geometric structure. The assumption in Conjecture 2.2 means that a geometric structure modelled on that of S is given on a Zariski open subset of X. By Theorem 2.1, the question is whether this structure is isomorphic to the standard one on S. There is a procedure in Cartan's theory to solve such problems. When a geometric structure is given, Cartan's procedure gives certain curvature tensors on some principal bundle whose vanishing will guarantee that the structure is locally isomorphic to the flat model of the given geometric structure. In fact, certain geometric structures modelled on rational homogeneous spaces have been studied in differential geometry ever since Cartan and the curvature tensors have been computed (e.g. [7], [30]). However, the model geometric structures on rational homogeneous spaces studied in these works do not necessarily agree with our geometric structure arising from the variety of minimal rational tangents. Recall that for each rational homogeneous space S of $b_2 = 1$, there is an associated simple root of the Lie algebra of automorphisms of S. When the rational homogeneous space S is associated to a long simple root, the two geometric structures coincide, as proved in [16]. These include Hermitian symmetric spaces and homogeneous contact manifolds, the latter meaning rational homogeneous spaces of $b_2 = 1$ where the isotropy representation on the tangent space has an irreducible subspace of codimension 1.

Thus at least for S associated to a long simple root, we have certain computational tools to check the validity of Conjecture 2.2: there are certain curvature tensors defined on the Zariski open subset of X at each point of which the variety of minimal rational tangents is isomorphic to the variety of minimal rational tangents at a point of S and the main problem is to show the vanishing of these curvature tensors. In some cases, these curvature tensors vanish for a purely local algebraic reason as proved in [31]. But in many cases, including the cases of Hermitian symmetric spaces and homogeneous contact manifolds, proving the vanishing of these curvature tensors requires a delicate geometric argument using minimal rational curves. One approach is to proceed in the following two steps. The first step is to prove the vanishing of the curvature assuming that the geometric structure modelled on S is defined on the whole X, in other words, that $\mathcal{C}_x$ is isomorphic to $\mathcal{C}'_s$ for each $x \in X$. Under this assumption, one studies the behavior of the curvature tensor along minimal rational curves to conclude the vanishing. This has been done for Hermitian symmetric spaces

in [12] and for homogeneous contact manifolds in [8]. The second step is to prove that the geometric structure modelled on S which is defined on a Zariski open subset of X extends to the whole X. This extension can be proved by looking at the local projective geometric invariants of the variety of minimal rational tangents $\mathcal{C}_x$ as x varies along a minimal rational curve, an idea introduced in [24]. In this manner, Conjecture 2.2 is proved for Hermitian symmetric spaces and homogeneous contact manifolds in [25]. It may be possible to extend these arguments further to prove Conjecture 2.2 when the rational homogeneous space is associated to a long simple root. However, Conjecture 2.2 for S associated to a short simple root looks much harder, and remains a challenge for future research.

Now that we know Conjecture 2.2 can be proved in many cases, how can we use this in the rigidity problems? The main question is how to get the assumption in Conjecture 2.2 from the conditions imposed on X in various rigidity questions. This depends on the individual rigidity problem and is often a difficult problem. We will see two examples below.

3. Deformation rigidity of rational homogeneous spaces

It is expected that the following generalization of Conjecture 3.5 in [28] holds. We will call it the deformation rigidity problem for rational homogeneous spaces.

Conjecture 3.1. Let $\pi : M \to \Delta$ be a smooth family of Fano manifolds such that the fiber M_t for each $t \neq 0$ is biholomorphic to a rational homogeneous space S. Then the central fiber M_0 is also biholomorphic to S.

More precisely speaking, this is the global deformation rigidity problem. It is well-known that the local deformation rigidity of rational homogeneous spaces follows from the vanishing $H^1(S, T(S)) = 0$, which is a consequence of Borel–Weil–Bott theorem.

Note that the Fano condition is necessary. In fact, the rational homogeneous space $S = \mathbb{P}_1 \times \mathbb{P}_1$ can be deformed to a non-Fano Hirzebruch surface. In this example, the automorphism group of S is not simple. One can also give examples whose automorphism groups are simple. For example, $S = \mathbb{P}T(\mathbb{P}_{2m+1})$ can be deformed to a non-Fano manifold of the form $\mathbb{P}(D \oplus L)$ where D is the null-correlation bundle and L is a line bundle on the odd-dimensional projective space $\mathbb{P}_{2m+1}$. In all these examples, $b_2(S) > 1$.

When $b_2(S) = 1$, the Fano condition is equivalent to the assumption that the central fiber is Kähler. In this case, it is expected that the following stronger rigidity holds.

Conjecture 3.2. Let $\pi : M \to \Delta$ be a smooth family of compact complex manifolds such that the fiber M_t for each $t \neq 0$ is biholomorphic to a rational homogeneous space S of $b_2 = 1$. Then the central fiber M_0 is also biholomorphic to S.

This is a generalization of Conjecture 3.4 in [28]. Conjecture 3.2 is proved for the projective space in [27] and for the hyperquadric in [9]. In these works, the main problem is to prove the existence of certain rational curves on M_0, which can play the role of minimal rational curves of Fano manifolds. Thus the nature of the problem is quite different from what we have discussed in Section 2. At the moment, it looks very hard to generalize the methods of [9] and [27] to other homogeneous spaces. For the other cases, the weaker conjecture, Conjecture 3.1, is already very interesting.

Regarding Conjecture 3.1, little work has been done when $b_2(S) > 1$. Conjecture 3.1 when $b_2(S) = 1$ has been proved in a series of works [10], [13] [16], [17], [19]. For the rest of the section, we will sketch the main ideas in this proof.

Taking the approach to rigidity problems discussed in Section 2, we see that one central problem is the following. Let $x \in M_0$ be a general point. Is the variety of minimal rational tangents at x, $\mathcal{C}_x \subset \mathbb{P}T_x(M_0)$, isomorphic to that of S?

To handle this question, we proceed as follows. Let us take a section $\{x_t, t \in \Delta\}$ of π such that $x_0 = x$. We compare the family of tangent morphisms

$$\tau_{x_t} \colon \mathcal{K}_{x_t} \to \mathcal{C}_{x_t} \subset \mathbb{P}T_{x_t}(M_t)$$

with the model

$$\tau_s \colon \mathcal{K}_s \to \mathcal{C}_s \subset \mathbb{P}T_s(S).$$

By the assumption τ_{x_t} is isomorphic to τ_s when $t \neq 0$. In particular, $\{\mathcal{K}_{x_t}, t \in \Delta\}$ is a smooth family of projective manifolds such that $\mathcal{K}_{x_t}$ is biholomorphic to $\mathcal{K}_s$ for $t \neq 0$. The first question to ask is whether $\mathcal{K}_{x_0}$ is biholomorphic to $\mathcal{K}_s$. This itself is a deformation rigidity problem. For many cases of S, $\mathcal{K}_s$ itself is a rational homogeneous space. In general, $\mathcal{K}_s$ is a variety very close to a rational homogeneous space. This indicates the possibility of using an induction on the dimension to solve Conjecture 3.1. Indeed, using additional information coming from the fact that $\mathcal{K}_{x_0}$ is the space of minimal rational curves through x_0, we can carry out this induction argument to show that $\mathcal{K}_{x_0}$ is biholomorphic to $\mathcal{K}_s$.

Now the problem is reduced to the study of the linear system defining τ_{x_t}. For the model S, τ_s is defined by a complete linear system when we view it as a morphism into the linear span of $\mathcal{C}_s$. Thus to show that τ_{x_0} is isomorphic to τ_s, it suffices to show that the linear span of $\mathcal{C}_{x_0}$ has the same dimension as that of $\mathcal{C}_s$. This leads to the study of the linear span of the variety of minimal rational tangents.

To crystalize the essential idea here, it is better to deal with a general Fano manifold X of $b_2 = 1$. Assume that we have a choice of the variety of minimal rational tangents $\mathcal{C} \subset \mathbb{P}T(X)$ such that for a general $x \in X$, $\mathcal{C}_x$ is irreducible. Let $D_x \subset T_x(X)$ be the linear span of $\mathcal{C}_x$. As x varies, D_x defines a Pfaffian system D on a Zariski open set of X. From the topological restriction $b_2(X) = 1$, one can show that D cannot be integrable. An essential property of D, which is one of the key points of [13], is that the Frobenius bracket at a general point

$$[\,,\,] \colon \Lambda^2 D_x \to T_x(X)/D_x$$

annihilates planes in D_x corresponding to tangent lines to the variety of minimal rational tangents $\mathcal{C}_x$. This follows from deformation theoretic properties of minimal rational curves.

Applying this general property of D to the variety of minimal rational tangents on M_0, one can show that if the linear span of $\mathcal{C}_{x_0}$ has dimension different from that of $\mathcal{C}_s$, then the Pfaffian system generated by $\mathcal{C}_{x_0}$ must be integrable, a contradiction to $b_2(M_0) = 1$. Thus we conclude that τ_{x_0} is isomorphic to τ_s.

From the discussion in Section 2, this already shows Conjecture 3.1 for Hermitian symmetric spaces and homogeneous contact manifolds. Also, when S is associated to a long simple root, we know, from Section 2, that there are some curvature tensors defined in a neighborhood of x_0 whose vanishing will guarantee that M_0 is biholomorphic to S. Since M_t has the same geometric structure modelled on S, the curvature tensor at x_0 is just the limit of the curvature tensor at x_t. As M_t, $t \neq 0$, is biholomorphic to S, the curvature tensor at x_t, $t \neq 0$, vanishes. By the continuity, the curvature tensor also vanishes at x_0. This completes the proof of Conjecture 3.1 when S is associated to a long simple root.

When S is associated to a short root, this argument does not work. In a sense, the difficulty lies in local differential geometry: the geometric structure given by the variety of minimal rational tangents has only part of the information needed for the geometric structure modelled on S studied by differential geometers. Thus we have only a 'crude' geometric structure, and in terms of this crude structure, S is not quite flat. One can still hope that there are certain structure constants involved and the constancy of the structure constants for $t \neq 0$ would imply that they remain unchanged at $t = 0$. Unfortunately, this idea can be worked out only in one special case [17].

For the final handling of Conjecture 3.1 for S associated to a short root, we do not deal with the equivalence problem directly. But instead we use the variety of minimal rational tangents to control the automorphism group of M_0. In a sense, instead of showing directly that the geometric structure defined by the variety of minimal rational tangents on M_0 is locally equivalent to that of S, we study the local automorphisms of the geometric structure to show that the group of automorphisms of M_0 is isomorphic to that of S. This suffices to conclude that M_0 is biholomorphic to S, in the setting of the deformation problem.

The study of the local automorphisms of a geometric structure leads to the theory of prolongations of linear Lie algebras (cf. [3], [21]). This theory is essentially algebraic and many works have been done in the context of the theory of filtered Lie algebras. Unfortunately, many results in this area requires that the linear Lie algebra involved is reductive, while the linear Lie algebra of the infinitesimal automorphisms of the variety of minimal rational tangents of a rational homogeneous space associated to a short root is not reductive. Thus from the view-point of Lie algebras, the variety of minimal rational tangents $\mathcal{C}_s$ is not really a nice object. But from the view-point of projective algebraic geometry, $\mathcal{C}_s$ is very nice: it is smooth and linearly normal, among other things. This motivates us to develop the theory of prolongations of linear

automorphisms of such nice projective varieties, using projective geometry instead of Lie algebra. In this theory, one makes use of the rich results on projective geometry such as [32] to replace the computational tool of semi-simple Lie algebras. This theory is developed in [19] to the extent needed for the proof of Conjecture 3.1 for S of $b_2 = 1$. As a byproduct, one can give new geometric proofs of many results on prolongations of reductive linear Lie algebras, too.

Readers must have noticed that our proof of Conjecture 3.1 for $b_2(S) = 1$ is not very uniform. A proof of Conjecture 2.2, or a further development of the prolongation theory of [19] would give a more uniform proof of Conjecture 3.1.

Conjecture 3.1 for the case of $b_2(S) > 1$ is wide open, except for some special cases. To attack this general situation, we would need to develop the theory of the variety of minimal rational tangents for Fano manifolds with $b_2 > 1$. Fano manifolds with $b_2 > 1$ admit non-trivial Mori contractions, which are crucial in understanding their geometry. Thus the geometry of the variety of minimal rational tangents has to be combined with Mori theory. This will be an interesting direction for future research.

4. The Campana–Peternell conjecture

The generalized Frankel conjecture which was proved by Mok in [23] states that a Fano manifold with a Kähler metric with nonnegative holomorphic bisectional curvature is a Hermitian symmetric space of compact type. In [1], Campana and Peternell proposed an algebraic generalization of this. Here we will discuss the following slightly stronger form of their conjecture:

Conjecture 4.1. A Fano manifold X is homogeneous if all rational curves on X are free.

A rational curve $\nu\colon \mathbb{P}_1 \to X$ is free if $\nu^*T(X)$ is a semi-positive vector bundle. There are many reasons to believe that Conjecture 4.1 is one of the central problems of the uniformization theory in several complex variables. In [5], it is proved that Conjecture 4.1 would give a complete description of Kähler manifolds with semi-positive curvature in a most reasonable sense.

One can check that Conjecture 4.1 is true for $\dim X \leq 3$ from the classification of Fano threefolds (cf. [1], [33]). However, in higher dimensions, very few results on Conjecture 4.1 are known. In the rest of the paper we will give a proof of the following which illustrates how the variety of minimal rational tangents can be used in rigidity problems.

Theorem 4.2. *Conjecture* 4.1 *is true in dimension* 4.

The proof uses the works of [2], [4], [22] and [24]. First of all, using Mori theory, [2] showed that Theorem 4.2 is true for Fano 4-folds with $b_2 > 1$. Thus we may

assume that $b_2(X) = 1$. Pick a family of minimal rational curves on X and consider the variety of minimal rational tangents $\mathcal{C}_x$ at a general point $x \in X$. Suppose $\dim \mathcal{C}_x = 3$. Then $X = \mathbb{P}_4$ by Theorem 2.3. Suppose $\dim \mathcal{C}_x = 2$. Then $X = \boldsymbol{Q}_3$ by Theorem 2.4. If $\dim \mathcal{C}_x = 0$, the freeness of all minimal rational curves implies $\mathcal{C}$ is an etale cover of X. Since a Fano manifold is simply connected, this shows that $\mathcal{C}$ is biholomorphic to X by the natural projection. This will give a regular foliation of X by minimal rational curves, a contradiction to $b_2(X) = 1$. Thus we are left with the case of $\dim \mathcal{C}_x = 1$. To handle this case, we use the following result of Mok from [24].

Theorem 4.3. *Let X be a Fano manifold with $b_2(X) = 1$. Assume that all rational curves on X are free and $\dim \mathcal{C}_x = 1$ for a general point $x \in X$. Assume in addition that $b_4(X) = 1$. Then X is homogeneous.*

Let us give a brief sketch of Mok's proof. By the freeness of rational curves, the space $\mathcal{K}$ of minimal rational curves is a projective manifold and the associated universal family morphisms $\rho\colon \mathcal{U} \to \mathcal{K}$, $\mu\colon \mathcal{U} \to X$ are smooth morphisms whose fibers are curves. By definition, ρ is a $\mathbb{P}_1$-bundle. If the fibers of μ have genus ≥ 1, μ must be a trivial fiber bundle over minimal rational curves on X. This contradicts some basic geometric feature of the morphism μ. Thus μ is a $\mathbb{P}_1$-bundle and the variety of minimal rational tangents at each $x \in X$ is a rational curve. The key point of [24] is to show that this rational curve $\mathcal{C}_x \subset \mathbb{P}T_x(X)$ has degree $d \leq 3$. Using this bound, one can see that the geometric structure defined by $\mathcal{C}$ is isomorphic to the one modelled on $\mathbb{P}_2$, $\boldsymbol{Q}_3$, or the 5-dimensional homogeneous contact manifold. Thus by the discussion in Section 2, X must be the model.

To get the bound on the degree d of $\mathcal{C}_x \subset \mathbb{P}T_x(X)$, Mok observed that there exists a stable vector bundle of rank 2 on $\mathcal{K}$ whose projectivization is $\rho\colon \mathcal{U} \to \mathcal{K}$. The crucial part of Mok's work is to show that the Chern number inequality for this stable bundle on $\mathcal{K}$ gives the bound $d \leq 3$. The additional assumption $b_4(X) = 1$ was used in this step. $b_4(X) = 1$ implies $b_4(\mathcal{K}) = 1$, which makes it possible to translate the Chern number inequality to the inequality $d \leq 3$.

In order to use Theorem 4.3 to complete the proof of Theorem 4.2, it suffices to show that the additional assumption $b_4(X) = 1$ in Theorem 4.3 can be removed. Since this part has not appeared in print, we will give full details.

First recall from [6] that the i-th Chow group $A_i(Z)$ of a variety Z is the abelian group of algebraic cycles of dimension i on Z modulo the rational equivalence. The i-th rational Chow group $A_i(Z)_{\mathbb{Q}}$ is the tensor product of $A_i(Z)$ with $\mathbb{Q}$. As mentioned above, the condition $b_4(X) = 1$ in the proof of Theorem 4.3 was used to get $b_4(\mathcal{K}) = 1$, which was necessary for the Chern class computation to get $d \leq 3$. Since Chern classes can be defined as algebraic cycles, it suffices to have $A_2(\mathcal{K})_{\mathbb{Q}} \cong \mathbb{Q}$ to conclude $d \leq 3$.

To prove $A_2(\mathcal{K})_{\mathbb{Q}} \cong \mathbb{Q}$ in the setting of Theorem 4.3 (without the assumption $b_4(X) = 1$), we use the double $\mathbb{P}_1$-bundle structure on $\mathcal{U}$ given by ρ and μ. We define, inductively, a sequence of smooth irreducible projective varieties Z_i, $i = 1, 2, 3 \ldots$,

of dimension i together with a morphism $\nu_i : Z_{i+1} \to Z_i$ and a morphism $\eta_i : Z_i \to \mathcal{U}$ as follows. Fix a general point $x \in X$. Let $Z_1 := \mu^{-1}(x)$ and $\eta_1 : Z_1 \to \mathcal{U}$ be the natural injection. Let Z_2 be the $\mathbb{P}_1$-bundle on Z_1 obtained as the pull-back of ρ by the morphism $\rho \circ \eta_1 : Z_1 \to \mathcal{K}$. Denote the bundle map $Z_2 \to Z_1$ by ν_1 and the natural map $Z_2 \to \mathcal{U}$ by η_2. Note that ν_1 has a natural section defined by η_1. Now let Z_3 be the $\mathbb{P}_1$-bundle over Z_2 obtained as the pull-back of μ by the morphism $\mu \circ \eta_2$. Denote the bundle map by $\nu_2 : Z_3 \to Z_2$ and the natural map $Z_3 \to \mathcal{U}$ by η_3. Then ν_2 has a natural section defined by η_2. Continuing in this manner, we define Z_{i+1} to be the $\mathbb{P}_1$-bundle over Z_i obtained as the pull-back of ρ or μ by $\rho \circ \eta_i$ or $\mu \circ \eta_i$ depending on whether i is odd or even. Denote the bundle map by $\nu_i : Z_{i+1} \to Z_i$ and the natural map to $\mathcal{U}$ by η_{i+1}. Then ν_i has a section defined by η_i.

Now we will apply the following lemma which is a simple consequence of Theorem 3.3 in [6].

Lemma 4.4. *Let $p : Z' \to Z$ be a $\mathbb{P}_1$-bundle with a section $\sigma : Z \to Z'$. Then any $\beta \in A_k(Z')$ is of the form $\beta = \sigma_*\alpha + p^*\gamma$ for some $\alpha \in A_k(Z)$ and $\gamma \in A_{k-1}(Z)$.*

Note that since X, $\mathcal{K}$ and Z_i are all rationally connected, $A_0(X)_{\mathbb{Q}} \cong A_0(\mathcal{K})_{\mathbb{Q}} \cong A_0(Z_i)_{\mathbb{Q}} \cong \mathbb{Q}$. Repeatedly applying Lemma 4.4, we see that $A_1(Z_i)$ is generated by curves which are sent by η_i to either a ρ-fiber or a μ-fiber in $\mathcal{U}$. Similarly, $A_2(Z_i)$ is generated by surfaces whose images in $\mathcal{U}$ under η_i are either a curve or surfaces of the form $\mu^{-1}(\mu(C))$ for some ρ-fiber C or $\rho^{-1}(\rho(C'))$ for some μ-fiber C'. It follows that the rank of the image of the push-forward

$$(\rho \circ \eta_i)_* : A_2(Z_i)_{\mathbb{Q}} \to A_2(\mathcal{K})_{\mathbb{Q}}$$

is ≤ 1. From $b_2(X) = 1$, there exists some ℓ such that $\eta_\ell : Z_\ell \to \mathcal{U}$ is surjective. Recall that when $\psi : Z \to Y$ is a proper surjective morphism of algebraic varieties, the induced push-forward map $\psi_* : A_i(Z)_{\mathbb{Q}} \to A_i(Y)_{\mathbb{Q}}$ is surjective. Thus

$$(\rho \circ \eta_\ell)_* : A_2(Z_\ell)_{\mathbb{Q}} \to A_2(\mathcal{K})_{\mathbb{Q}}$$

is surjective. It follows that $A_2(\mathcal{K})_{\mathbb{Q}} \cong \mathbb{Q}$. This finishes the proof of Theorem 4.2.

References

[1] Campana, F., Peternell, T., Projective manifolds whose tangent bundles are numerically effective. *Math. Ann.* **289** (1991), 169–187.

[2] Campana, F., Peternell, T. 4-folds with numerically effective tangent bundles and second Betti numbers greater than one. *Manuscripta Math.* **79** (1993), 225–238.

[3] Cartan, E., Les groupes de transformations continus, infinis, simples. *Ann. Sci. École Norm. Sup.* **26** (1909), 93–161.

[4] Cho, K., Miyaoka, Y., Shepherd-Barron, N., Characterizations of projective space and applications to complex symplectic manifolds. In *Higher dimensional birational geometry* (Kyoto, 1997), Adv. Stud. Pure Math. 35, Math. Soc. Japan, Tokyo 2002, 1–88.

[5] Demailly, J.-P., Peternell, T., Schneider, M., Compact complex manifolds with numerically effective tangent bundles. *J. Algebraic Geom.* **3** (1994), 295–345.

[6] Fulton, W., *Intersection theory*. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) 2, Springer-Verlag, Berlin 1984.

[7] Guillemin, V., The integrability problem for G-structures. *Trans. Amer. Math. Soc.* **116** (1965), 544–560.

[8] Hong, J., Fano manifolds with geometric structures modeled after homogeneous contact manifolds. *International J. Math.* **11** (2000), 1203–1230.

[9] Hwang, J.-M., Nondeformability of the complex hyperquadric. *Invent. Math.* **120** (1995), 317–338.

[10] Hwang J.-M., Rigidity of homogeneous contact manifolds under Fano deformation. *J. Reine Angew. Math.* **486** (1997), 153–163.

[11] Hwang, J.-M., Geometry of minimal rational curves on Fano manifolds. In *School on Vanishing Theorems and Effective Results in Algebraic Geometry* (Trieste, 2000), ICTP Lect. Notes 6, Abdus Salam International Centre for Theoretical Physics, Trieste 2001, 335–393.

[12] Hwang, J.-M., Mok, N., Uniruled projective manifolds with irreducible reductive G-structures. *J. Reine Angew. Math.* **490** (1997), 55–64.

[13] Hwang J.-M., Mok N., Rigidity of irreducible Hermitian symmetric spaces of the compact type under Kähler deformation. *Invent. Math.* **131** (1998), 393–418.

[14] Hwang J.-M., Mok N., Varieties of minimal rational tangents on uniruled projective manifolds. In *Several Complex Variables* (ed. by Michael Schneider, Yum-Tong Siu), Math. Sci. Res. Inst. Publ. 37, Cambridge University Press, Cambridge 1999, 351–389.

[15] Hwang, J.-M., Mok, N., Cartan-Fubini type extension of holomorphic maps for Fano manifolds with Picard number 1. *J. Math. Pures Appl.* (9) **80** (2001), 563–575.

[16] Hwang, J.-M., Mok, N., Deformation rigidity of the rational homogeneous space associated to a long simple root. *Ann. Sci. École Norm. Sup.* (4) **35** (2002), 173–184.

[17] Hwang, J.-M., Mok, N., Deformation rigidity of the 20-dimensional F_4-homogeneous space associated to a short root. In *Algebraic Transformation Groups and Algebraic Varieties* (ed. by V. L. Popov), Encyclopaedia Math. Sci. 132, Springer-Verlag, Berlin 2004, 37–58.

[18] Hwang, J.-M., Mok, N., Birationality of the tangent map for minimal rational curves. *Asian J. Math.* **8** (Special issue dedicated to Y.-T. Siu on his 60th birthday) (2004), 51–64.

[19] Hwang, J.-M., Mok, N., Prolongations of infinitesimal linear automorphisms of projective varieties and rigidity of rational homogeneous spaces of Picard number 1 under Kähler deformation. *Invent. Math.* **160** (2005), 591–645.

[20] Kebekus, S., Families of singular rational curves. *J. Alg. Geom.* **11** (2002), 245–256.

[21] Kobayashi, S. *Transformation groups in differential geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete 70, Springer-Verlag, New York, Heidelberg 1972.

[22] Miyaoka, Y., Numerical characterisations of hyperquadrics. In *Complex analysis in several variables—Memorial Conference of Kiyoshi Oka's Centennial Birthday*, Adv. Stud. Pure Math. 42, Math. Soc. Japan, Tokyo 2004, 209–235.

[23] Mok, N., The uniformization theorem for compact Kähler manifolds of nonnegative holomorphic bisectional curvature. *J. Differential Geom.* **27** (1988), 179–214.

[24] Mok, N., On Fano manifolds with nef tangent bundles admitting 1-dimensional varieties of minimal rational tangents. *Trans. Amer. Math. Soc.* **354** (2002), 2639–2658.

[25] Mok, N., Recognizing certain rational homogeneous manifolds of Picard number 1 from their varieties of minimal rational tangents. Preprint.

[26] Mori, S., Projective manifolds with ample tangent bundles. *Ann. Math.* **110** (1979), 593–606.

[27] Siu, Y.-T., Nondeformability of the complex projective space. *J. Reine Angew. Math.* **399** (1989), 208–219; Errata. *J. Reine Angew. Math.* **431** (1992), 65–74.

[28] Siu, Y.-T., Uniformization in several complex variables. In *Contemporary Geometry: J.-Q Zhong memorial volume* (ed. by Hung-Hsi Wu). Univ. Ser. Math., Plenum, New York 1991, 95–130.

[29] Siu, Y.-T., Yau, S.-T., Compact Kähler manifolds of positive bisectional curvature. *Invent. Math.* **59** (1980), 189–204.

[30] Tanaka, N., On the equivalence problems associated with simple graded Lie algebras. *Hokkaido Math. J.* **8** (1979), 23–84.

[31] Yamaguchi, K., Differential systems associated with simple graded Lie algebras. In *Progress in Differential Geometry*, Adv. Stud. Pure Math. 22, Math. Soc. Japan, Tokyo 1993, 413–494.

[32] Zak, F. L., *Tangents and Secants of Algebraic Varieties*. Transl. Math. Monographs 127, Amer. Math. Soc., Providence, RI, 1993.

[33] Zheng, F., On semi-positive threefolds. Thesis, Harvard, 1990.

School of Mathematics, Korea Institute for Advanced Study, Seoul 130-722, Korea
E-mail: jmhwang@kias.re.kr

Geometry of multiple zeta values

Tomohide Terasoma

Abstract. Many relations are known between multiple zeta values $\zeta(k_1, \dots, k_n)$. A relation coming from the associator condition for the Drinfeld associator, the generating function of multiple zeta values, is a geometric relation. By the theory of mixed motives, we can control the dimension of the rational linear hull of multiple zeta values. The harmonic shuffle relation also comes from geometry, and more strongly, this is implied by the associator relation.

Mathematics Subject Classification (2000). Primary 14C30; Secondary 14F42.

Keywords. Multiple zeta value, mixed motif.

1. Introduction

Let $k_1, \dots, k_n \geq 1$ be integers such that $k_n \geq 2$. We define a multiple zeta value $\zeta(k_1, \dots, k_n)$ by

$$\zeta(k_1, \dots, k_n) = \sum_{0<m_1<m_2<\cdots<m_n} \frac{1}{m_1^{k_1} \cdots m_n^{k_n}}.$$

We define the weight w of the index $(k_1, \dots, k_n)$ by $w = k_1 + \cdots + k_n$. Many relations between multiple zeta values are known for a long time; for example,

$$\zeta(2) \cdot \zeta(2) = 4\zeta(1,3) + 2\zeta(2,2), \quad \zeta(3) = \zeta(1,2), \quad \zeta(4) = 4\zeta(1,3).$$

The first one is a quadratic relation and the second and the third ones are linear relations between multiple zeta values. Several systematic methods are known to produce a series of relations for multiple zeta values: iterated integral shuffle relation, duality relation, harmonic shuffle relation, and so on. The iterated integral relation and the duality relation are a part of the associator relation, which is closely related to the Grothendieck–Teichmüller group. These relations produce many linear relations between multiple zeta values. What is very interesting is that all the known rational relations preserve the weights introduced above. So it is natural to expect that all $\mathbb{Q}$-relations come from geometry.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. Iterated integral expression

A multiple zeta value has an iterated integral expression, which enables us to study multiple zeta values from a geometric point of view. Let $\omega_1, \dots, \omega_n$ be one-forms on a manifold X and let $\gamma\colon [0,1] \to X$ be a path starting from a point a and ending with a point b. An iterated integral is defined by

$$\int_\gamma \omega_1\omega_2\cdots\omega_n = \int_{0<t_n<t_{n-1}<\cdots<t_1<1} \mathrm{pr}_1^*\,\omega_1 \wedge \mathrm{pr}_2^*\,\omega_2 \wedge \cdots \wedge \mathrm{pr}_n^*\,\omega_n,$$

where $\mathrm{pr}_i\colon [0,1]^n \to [0,1]$ is the i-th projection. A multiple zeta value is expressed as

$$\zeta(k_1,\dots,k_n) = \int_{[0,1]} \left(\frac{dx}{x}\right)^{k_n-1} \frac{dx}{1-x} \cdots \left(\frac{dx}{x}\right)^{k_1-1} \frac{dx}{1-x}.$$

To control many relations it is convenient to consider the "generating function" of multiple zeta values which is called the Drinfeld associator ([Dr]). Let $A = \mathbb{C}\langle\langle e_0, e_1\rangle\rangle$ be the non-commutative formal power series ring generated by e_0 and e_1 over $\mathbb{C}$ and $\omega = \frac{e_0 dx}{x} + \frac{e_1 dx}{x-1}$ an A-valued one-form. The Drinfeld associator Φ is defined by

$$\Phi = \Phi(e_0, e_1) = \lim_{t\to 0} t^{-e_1}\left[\exp\int_t^{1-t}\omega\right]t^{e_0},$$

where $\exp\int_a^b \omega = 1 + \sum_{i=1}^\infty \int_a^b \underbrace{\omega\cdots\omega}_{i\text{-times}}$. The multiple zeta value $(-1)^n\zeta(k_1,\dots,k_n)$ appears as the coefficient of $e_0^{k_n-1}e_1\cdots e_0^{k_1-1}e_1$.

3. Period of fundamental group of $\mathbb{P}^1 - \{0, 1, \infty\}$

The Drinfeld associator Φ describes the period of the fundamental group of $\mathcal{M}_{0,4} = \mathbb{P}^1 - \{0,1,\infty\}$ for tangential base points introduced by Deligne [De]. Let $|\mathcal{M}_{0,4}| = \{\vec{01}, \vec{10}, \dots\}$ be a set of tangential points of $\mathcal{M}_{0,4}$ and p, q be elements of $|\mathcal{M}_{0,4}|$. The comparison isomorphism

$$\mathrm{comp}\colon \mathbb{Q}[\pi_1(\mathbb{P}^1 - \{0,1,\infty\}, p, q)]\hat{} \otimes \mathbb{C} \simeq \mathbb{Q}\langle\langle e_0, e_1\rangle\rangle \otimes \mathbb{C} \tag{1}$$

defines a mixed Hodge structure on $\mathbb{Q}[\pi_1(\mathbb{P}^1 - \{0,1,\infty\}, p, q)]\hat{}$. The Drinfeld associator is equal to the image of $[0,1]$ under the comparison homomorphism. The Drinfeld associator satisfies the associator relations arising from the compatibility condition for homomorphisms of mixed Hodge structures. In general, the associator relations can be formulated via the compatibility condition for isomorphisms between topological and de Rham fundamental groups. To interpret an associator as a functorial isomorphism we introduce the fundamental category (Fund) and the notion of algebroids.

Definition 3.1. Let K be a field. A K-algebroid $\mathcal{U} = \{\mathcal{U}_{ab}\}_{ab}$ over a set B consists of

(1) a family of K-vector spaces $\{\mathcal{U}_{ab}\}_{ab}$ indexed by $a, b \in B$, and
(2) a family of homomorphisms

$$\mathcal{U}_{bc} \otimes \mathcal{U}_{ab} \to \mathcal{U}_{ac},$$

which is associative.

We assume the following properties:

(1) For $a \in B$ the associative ring $\mathcal{U}_{aa}$ has a unit.
(2) The vector space $\mathcal{U}_{ab}$ is a free left $\mathcal{U}_{bb}$-module of rank one and a free right $\mathcal{U}_{aa}$-module of rank one.
(3) The natural homomorphism

$$\mathcal{U}_{bc} \otimes_{\mathcal{U}_{bb}} \mathcal{U}_{ab} \to \mathcal{U}_{ac}$$

is an isomorphism.

We can define the notion of K-Hopf algebroids similarly. For a $\mathbb{Q}$-algebraic variety, by attaching a completion of the $\mathbb{Q}$-linear hull of Betti and de Rham fundamental groupoids, we get two functors

$$\mathcal{U}^{\mathrm{B}}, \mathcal{U}^{\mathrm{DR}} \colon (\mathrm{Var}/\mathbb{Q}) \to (\text{Hopf-alg}/\mathbb{Q})$$

from the category $(\mathrm{Var}/\mathbb{Q})$ of $\mathbb{Q}$-algebraic varieties to the category $(\text{Hopf-alg}/\mathbb{Q})$ of $\mathbb{Q}$-Hopf algebroids. The comparison map obtained by Hodge theory gives a functorial isomorphism of these two functors over $\mathbb{C}$.

The object of the fundamental category (Fund) consists of four spaces $\mathcal{M}_{0,4}$, $\mathcal{M}_{0,5}$, $\overline{\mathcal{M}_{0,4}} - \{0, \infty\}$ and the punctured disc Δ^*, with tangential points. Morphisms are generated by

(1) inclusions $\Delta^* \to \mathcal{M}_{0,4}$ around points 0, 1 or ∞,
(2) "infinitesimal inclusions" $\mathcal{M}_{0,4} \to \mathcal{M}_{0,5}$, and
(3) natural inclusion $\mathcal{M}_{0,4} \to \overline{\mathcal{M}_{0,4}} - \{0, \infty\}$.

We can define two functors of Betti and de Rham fundamental algebroids over a set of tangential points $\mathcal{U}^{\mathrm{B}}(?) = \mathbb{Q}[\pi_1^{\mathrm{B}}(?)]\hat{}$ and $\mathcal{U}^{\mathrm{DR}}(?) = \mathbb{Q}[\pi_1^{\mathrm{DR}}(?)]\hat{}$: $(\mathrm{Fund}) \to (\text{Hopf-alg}/\mathbb{Q})$. The set of functorial isomorphisms from $\mathcal{U}^{\mathrm{B}} \otimes \mathbb{C}$ to $\mathcal{U}^{\mathrm{DR}} \otimes \mathbb{C}$ is denoted by $\mathrm{Isom}(\mathcal{U}^{\mathrm{B}} \otimes \mathbb{C}, \mathcal{U}^{\mathrm{DR}} \otimes \mathbb{C})$.

Definition 3.2 (Associator). (1) Let ρ and e be the canonical generators of $\pi_1(\Delta^*, +)$ and the dual of $\frac{dx}{x}$ in $\mathcal{U}^{\mathrm{DR}}(\Delta^*)$. Here $+$ denotes a tangential point of Δ^* defined by the local coordinate x. For an element $\varphi \in \mathrm{Isom}(\mathcal{U}^{\mathrm{B}} \otimes \mathbb{C}, \mathcal{U}^{\mathrm{DR}} \otimes \mathbb{C})$ we define $\lambda(\varphi) \in \mathbb{C}^\times$ by $\lambda(\varphi) = \varphi(\log(\rho))/e$. Thus we have a map

$$\lambda \colon \mathrm{Isom}(\mathcal{U}^{\mathrm{B}} \otimes \mathbb{C}, \mathcal{U}^{\mathrm{DR}} \otimes \mathbb{C}) \to \mathbb{C}^\times.$$

(2) We define the set of associator Ass as the inverse image

$$\lambda^{-1}(2\pi i) \subset \mathrm{Isom}(\mathcal{U}^{\mathrm{B}} \otimes \mathbb{C}, \mathcal{U}^{\mathrm{B}} \otimes \mathbb{C})$$

of $2\pi i$ under the map λ.

Remark 3.3. The functorial isomorphism φ is determined by the element $\Phi = \varphi([0,1]) \in \mathcal{U}^{\mathrm{DR}}(\mathcal{M}_{0,4}) = \mathbb{C}\langle\langle e_0, e_1\rangle\rangle$. The condition for an element Φ to be able to be continued to a functorial isomorphism is nothing but the classical condition for associators (see [Dr]). By this correspondence, the Drinfeld associator corresponds to the classical comparison map. This has been communicated to the author by M. Matsumoto.

Definition 3.4 (Grothendieck–Teichmüller group). (1) Let X be an algebraic variety over $\mathbb{Q}$. The $\mathbb{Q}_l$ completion of the etale fundamental groupoid of a variety $X \otimes \overline{\mathbb{Q}}$ is denoted by $\pi_1^l(X \otimes \overline{\mathbb{Q}})$. The completion of the $\mathbb{Q}_l$-linear hull of $\pi_1^l(X \otimes \overline{\mathbb{Q}})$ is denoted by $\mathcal{U}^l(X)$. If X is one of $\Delta^* = \mathrm{Spec}(\mathbb{Q}[\![x]\!])$, $\mathcal{M}_{0,4}$ or $\mathcal{M}_{0,5}$, we can similarly define similar tangential base points (see [I]). The Hopf algebroid $\mathcal{U}^l(X)$ gives a functor $\mathcal{U}^l$: (Fund) $\to$ (Hopf-alg/$\mathbb{Q}_l$).

(2) Let $* = \mathrm{B}, \mathrm{DR}$ or l. The group $\mathrm{GT}_* = \mathrm{Aut}(\mathcal{U}^*)$ of functorial automorphisms of $\mathcal{U}^*$ is called the $*$-Grothendieck–Teichmüller group. The groups GT_{B}, $\mathrm{GT}_{\mathrm{DR}}$ and GT_l are pro-algebraic groups over $\mathbb{Q}$, $\mathbb{Q}$ and $\mathbb{Q}_l$, respectively. For an element of $\varphi \in \mathrm{Aut}(\mathcal{U}^*)$, by attaching $\varphi(\Delta^*) \in \mathrm{Aut}_{\text{Hopf-alg}}(\mathcal{U}^*(\Delta^*))$, we have a homomorphism of pro-algebraic groups:

$$\lambda\colon\ \mathrm{GT}_* = \mathrm{Aut}(\mathcal{U}^*) \to \boldsymbol{G}_m,$$

where $\boldsymbol{G}_m$ is the multiplicative group. The kernel of $\lambda\colon\ \mathrm{GT}_* \to \boldsymbol{G}_m$ is denoted by $\mathrm{GT}_*^{(1)}$. It is a pro-nilpotent algebraic group.

(3) By the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ we have a natural homomorphism $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GT}_l$.

4. Mixed Tate motives and Grothendieck–Teichmüller group

Following Levine, Voevodsky, Hanamura ([L]) we can define an abelian category $\mathrm{MTM}_{\mathbb{Q}}$ of mixed Tate motives over $\mathbb{Q}$. Goncharov ([G]) defined a full subcategory $\mathrm{MTM}_{\mathbb{Z}}$ of mixed Tate motives over $\mathbb{Z}$. By the classical comparison map (1), $\mathcal{U}^{\mathrm{B+DR}}(\mathbb{P}^1 - \{0,1,\infty\})_{\vec{01},\vec{10}}$ becomes a mixed Hodge structure of mixed Tate type. This mixed Hodge structure is obtained by the nearby fiber (= the limit of mixed Hodge structure) at $\vec{01}, \vec{10}$ of the variation of mixed Hodge structure $\mathcal{U}(\mathbb{P}^1 - \{0,1,\infty\})_{ab}$ of two variables a, b. It is natural to expect that the near by fiber of a family of mixed motives is also a mixed motif, which is not well formulated up to now. In fact, for Hodge realization, a limit of a mixed Hodge structure depends on "the tangential structure", i.e. the choice of a branch of "$\log(t)$". But in our setting, we are very happy to have the following.

Theorem 4.1 (Deligne–Goncharov [DG], Terasoma [T]). *There exists an object* $\mathcal{U}^M(\mathbb{P}^1 - \{0, 1, \infty\})^{\rightarrow\ \rightarrow}_{01,10}$ *in* $\mathrm{MTM}_{\mathbb{Z}}$ *whose Hodge realization is isomorphic to* $\mathcal{U}^{\mathrm{B+DR}}(\mathbb{P}^1 - \{0, 1, \infty\})^{\rightarrow\ \rightarrow}_{01,10}$.

We state a consequence of the above theorem in the language of Tannakian category. Let $H^{\mathrm{DR}}\colon \mathrm{MTM}_{\mathbb{Z}} \to (\mathrm{Vec}_{\mathbb{Q}})$ be the de Rham realization functor, let $\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}})$ be the Tannaka fundamental group (see [DM]) of $\mathrm{MTM}_{\mathbb{Z}}$ for the fiber functor H^{DR} and let $\lambda\colon \pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}}) \to \boldsymbol{G}_m$ be the homomorphism obtained by the functor attaching the associated graded module for weight filtration. The kernel of λ is denoted by $\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}})^{(1)}$. Let $\mathcal{U}(\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}}))$ be the completion of the $\mathbb{Q}$-linear hull of $\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}})$. An object in $\mathrm{MTM}_{\mathbb{Z}}$ corresponding to the representation $\mathcal{U}(\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}}))$ of $\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}})$ is denoted by $\mathcal{U}^M(\pi_1(\mathrm{MTM}_{\mathbb{Z}}))$. By the above theorem and the definition of Tannaka fundamental group, $H^{\mathrm{DR}}\mathcal{U}^M(\mathbb{P}^1 - \{0, 1, \infty\})^{\rightarrow\ \rightarrow}_{01,10})$ becomes a (homogeneous) $\mathcal{U}(\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{DR}}))$-module. Therefore the periods of $\mathcal{U}^M(\mathbb{P}^1 - \{0, 1, \infty\})$ are controlled by those of $\mathcal{U}^M(\pi_1(\mathrm{MTM}_{\mathbb{Z}}))$. Taking into account the real structures we have the following corollary.

Corollary 4.2 (Conjectured by Zagier, ([G], [T])). *Let* L_n *be the* $\mathbb{Q}$*-vector space generated by multiple zeta values of weight* n*. Then we have* $\dim_{\mathbb{Q}} L_n \leq d_n$*, where* d_n *is defined by the generating function* $\sum_{i=0}^{\infty} d_n t^n = \frac{1}{1-t^2-t^3}$.

Either

(1) by comparing $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and $\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^l)$ via the homomorphism $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^l)$ by using Hain–Matsumoto's result, or

(2) by using the infinitesimal embedding $\mathcal{U}^M(\mathcal{M}_{0,4}) \to \mathcal{U}^M(\mathcal{M}_{0,5})$ of mixed Tate motives according to Deligne–Goncharov,

we have a homomorphism of pro-algebraic groups

$$\mathrm{Rep}\colon \pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{B}}) \to \mathrm{GT}_{\mathrm{B}}$$

which is compatible with the weight homomorphisms λ. The Deligne–Ihara conjecture asserts

Conjecture 4.3 (Deligne–Ihara). The homomorphism Rep is an isomorphism.

Remark 4.4. The injectivity of Rep is equivalent to the following statement:

Any mixed Tate motif over $\mathbb{Z}$ is generated by $\mathcal{U}^M(\mathbb{P}^1 - \{0, 1, \infty\})^{\rightarrow\ \rightarrow}_{01,10}$.

As a consequence the injectivity of Rep implies that periods of any mixed Tate motif over $\mathbb{Z}$ are $\mathbb{Q}$-linear combinations of multiple zeta values.

Note that $\mathcal{U}(\pi_1(\mathrm{MTM}_{\mathbb{Z}}, H^{\mathrm{B}})^{(1)})$ is known to be a free Lie algebra generated by $c_3, c_5, c_7, \ldots$ ([DG]), and the description of GT_{B} is combinatorial. Therefore the above conjecture is (in principle) a combinatorial question. It is answered in the positive up to a quite high degree.

5. Harmonic shuffle relation

In this section we introduce the harmonic shuffle relation according to Hoffman [H] and its regularized version by Ihara–Kaneko–Zagier and Racinet ([IKZ], [R]). The rearrangement of the product of two multiple zeta values expressed by infinite series leads to a $\mathbb{Z}$-linear combination of multiple zeta values: for example, we have

$$\zeta(2)\zeta(2) = \sum_{0<n} \frac{1}{n^2} \cdot \sum_{0<m} \frac{1}{m^2} = \sum_{0<n} \frac{1}{n^4} + 2 \sum_{0<n<m} \frac{1}{n^2 m^2} = \zeta(4) + 2\zeta(2,2).$$

A relation of this type is called harmonic shuffle relation. Using the technique of regularization we also consider a relation of this type for non-convergent sums. As a consequence Racinet and Ihara–Kaneko–Zagier obtained a wider class of relations, the so-called "regularized harmonic shuffle relation". We briefly recall the formulation by Racinet of the regularized harmonic shuffle relation. We have the following approximation of the partial summation

$$\begin{aligned} \zeta_N(k_1, \dots, k_n) &= \sum_{0<m_1<m_2<\cdots<m_n<N} \frac{1}{m_1^{k_1} \dots m_n^{k_n}} \\ &= P_{k_1,\dots,k_n}(\log N + \gamma) + O(N^{-\epsilon}) \end{aligned} \tag{2}$$

by a real coefficient polynomial $P_{k_1,\dots,k_n}(T)$. Using Boutet de Monvel–Zagier's theorem, the generating series of $P_{k_1,\dots,k_n}(T)$ can be computed from the Drinfeld associator. By rearranging the partial summation up to N and using the asymptotic expression (2), the product $P_{k_1,\dots,k_n}(T)P_{k'_1,\dots,k'_{n'}}(T)$ becomes a $\mathbb{Z}$-linear combination of polynomials of the form $P_{k''_1,\dots,k''_{n''}}(T)$. In order to formulate the result it is convenient to introduce a new coproduct, the harmonic coproduct. Let W be a subalgebra $\mathbb{C} \oplus \mathbb{C}\langle\langle e_0, e_1\rangle\rangle e_1$ of $\mathbb{C}\langle\langle e_0, e_1\rangle\rangle$. This algebra is isomorphic to a weighted completion $\mathbb{C}\langle\langle y_1, y_2 \dots\rangle\rangle$ of the free algebra generated by $y_1 = -e_1$, $y_2 = -e_0e_1, \dots$, $y_i = -e_0^{i-1}e_1, \dots$. We define a harmonic coproduct $\Delta_*\colon W \to W \otimes W$ by an algebra homomorphism given by

$$\Delta_*(y_n) = \sum_{i=0}^{n} y_i \otimes y_{n-i},$$

with $y_0 = 1$. Let $\Phi_{\mathrm{DR}} = 1 + \varphi_0 e_0 + \varphi_1 e_1$ be the Drinfeld associator and set $\Phi_{\mathrm{DR},Y} = 1 + \varphi_1 e_1 \in W$. Then there is a unique formal power series $\Gamma_{\mathrm{DR}}(s) \in 1 + s^2\mathbb{C}[\![s]\!]$ such that

$$\Phi^{\mathrm{ab}}_{\mathrm{DR},Y}(e_0, e_1) = \frac{\Gamma_{\mathrm{DR}}(-e_0)\Gamma_{\mathrm{DR}}(-e_1)}{\Gamma_{\mathrm{DR}}(-e_0 - e_1)},$$

where $\Phi^{\mathrm{ab}}_{\mathrm{DR},Y}$ is the image of $\Phi_{\mathrm{DR},Y}$ in the abelianization $\mathbb{C}[\![e_0, e_1]\!]$. We define the modified Y-series $\Phi^{\mathrm{mod}}_{\mathrm{DR},Y}$ by $\Phi^{\mathrm{mod}}_{\mathrm{DR},Y} = \Gamma_{\mathrm{DR}}(y_1)^{-1}\Phi_{\mathrm{DR},Y} \in W$.

Theorem 5.1 (Racinet [R], Ihara–Kaneko–Zagier [IKZ] in a different formalism). *With the above notation, we have*

$$\Delta_*(\Phi_{\mathrm{DR},Y}^{\mathrm{mod}}) = \Phi_{\mathrm{DR},Y}^{\mathrm{mod}} \otimes \Phi_{\mathrm{DR},Y}^{\mathrm{mod}}.$$

In other words, $\Phi_{\mathrm{DR},Y}^{\mathrm{mod}}$ *is a group-like element under the harmonic coproduct.*

6. Fake Hodge realization and harmonic shuffle relation

As is shown in Sections 3 and 5, the origin of the harmonic shuffle relation comes from the rearrangement of partial series, and that of the associator relation comes essentially from the functoriality for infinitesimal inclusions from $\mathcal{M}_{0,4}$ into $\mathcal{M}_{0,5}$. Though two origins are quite different, we can show that the associator relation implies the harmonic shuffle relation. The contents of this section is a result of collaboration with P. Deligne.

For an associator $\Phi = 1+\varphi_0 e_0+\varphi_1 e_1$ set $\Phi_Y = 1+\varphi_1 e_1 = \Phi_Y(y_1, y_2, \dots) \in W$.

Theorem 6.1 (Deligne–Terasoma). (1) *Let* Φ_Y^{ab} *be the image of* Φ_Y *in the abelianization* $\mathbb{C}[\![e_0, e_1]\!]$. *Then there exists a unique element* $\Gamma_\Phi(s)$ *in* $1 + s^2\mathbb{C}[\![s]\!]$ *such that*

$$\Phi_Y^{\mathrm{ab}}(e_0, e_1) = \frac{\Gamma_\Phi(-e_0)\Gamma_\Phi(-e_1)}{\Gamma_\Phi(-e_0 - e_1)}.$$

(2) *Using* Γ_Φ *obtained in* (1) *we define*

$$\Phi_Y^{\mathrm{mod}}(y_1, y_2, \dots) := \Gamma_\Phi(y_1)^{-1}\Phi_Y(y_1, y_2, \dots) \in W.$$

Then we have

$$\Delta_*(\Phi_Y^{\mathrm{mod}}) = \Phi_Y^{\mathrm{mod}} \otimes \Phi_Y^{\mathrm{mod}}.$$

The proof of the above theorem is divided into two parts.

6.1. Fake Hodge realization attached to an associator. By Definition 3.2 an associator defines Hopf algebroid objects $\mathcal{U}(\mathcal{M}_{0,4})$ and $\mathcal{U}(\mathcal{M}_{0,5})$ in the category

$$\begin{aligned}\mathcal{M} &= \mathrm{Vec}_{\mathbb{Q}} \times_{\mathrm{Vec}_{\mathbb{C}}} \mathrm{Vec}_{\mathbb{Q}} \\ &= \{(V^{\mathrm{B}}, V^{\mathrm{DR}}, \varphi) \mid V^{\mathrm{B}} \text{ and } V^{\mathrm{DR}} \text{ are } \mathbb{Q}\text{-vector spaces, } \varphi\colon V^{\mathrm{B}} \otimes \mathbb{C} \xrightarrow{\simeq} V^{\mathrm{DR}} \otimes \mathbb{C} \\ &\qquad\qquad \text{is an isomorphism of } \mathbb{C} \text{ vector spaces}\}.\end{aligned}$$

The i-th projection induces a homomorphism $\mathrm{pr}_i\colon \mathcal{U}(\mathcal{M}_{0,5}) \to \mathcal{U}(\mathcal{M}_{0,4})$ of Hopf algebroids in $\mathcal{M}$. For an algebroid $\mathcal{U}$ in $\mathcal{M}$, we can introduce a notion of "$\mathcal{U}$-modules" in the category $\mathcal{M}$. For a $\mathcal{U}(\mathcal{M}_{0,5})$-module M we can define a $\mathcal{U}(\mathcal{M}_{0,4})$-module $R^j \mathrm{pr}_i M$ by using cohomological technology. Note that an isomorphism

$$R^j \mathrm{pr}_i M^{\mathrm{B}} \otimes \mathbb{C} \xrightarrow{\simeq} R^j \mathrm{pr}_i M^{\mathrm{DR}} \otimes \mathbb{C}$$

is encoded in the object $R^j \operatorname{pr}_i M$, which can be computed from the associator chosen first. This isomorphism is called a fake comparison map associated to an associator. We can also define the notion of "perverse" sheaf, and so on. These modules are equipped with a mixed Hodge structure which is different from the natural one. They are called a fake Hodge realization.

6.2. Multiplicative convolution. Following Deligne we introduce a geometric interpretation of harmonic shuffle relation. Let $\mathcal{A}$ be the category of topological perverse sheaves on $\overline{\mathcal{M}_{0,4}} - \{0, \infty\} \simeq \boldsymbol{G}_m$ smooth outside of $\{1\}$ with a nilpotent monodromy. We introduce an equivalence relation on $\mathcal{A}$ as follows. A morphism $\mathcal{F}_1 \to \mathcal{F}_2$ in $\mathcal{A}$ is equivalent if the induced homomorphism of vanishing cycles is an isomorphism. The quotient category under this equivalence relation is denoted by $\bar{\mathcal{A}}$. Then $\bar{\mathcal{A}}$ is again an abelian category and we can show that this abelian category is equivalent to the category of W^{B} modules, where

$$W^{\mathrm{B}} = \mathcal{U}^{\mathrm{B}}_{\substack{\to \\ 10,10}} \log \rho_1 \oplus \mathbb{Q},$$

ρ_1 being the canonical generator of local monodromy around $\{1\}$. For two perverse sheaves $\mathcal{F}_1$ and $\mathcal{F}_2$, we can define a biadditive functor $*\colon \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ by the multiplicative convolution

$$\mathcal{F}_1 * \mathcal{F}_2 = {}^p\mathcal{H}^1 \mathbb{R} \operatorname{pr}_{5*}(\operatorname{pr}_1^*(\mathcal{F}_1) \otimes \operatorname{pr}_2^*(\mathcal{F}_2)).$$

The equivalence class of the above convolution $\mathcal{F}_1 * \mathcal{F}_2$ depends only on the equivalence class of $\mathcal{F}_1$ and $\mathcal{F}_2$ and, as a consequence, the convolution induces a biadditive functor $*\colon \bar{\mathcal{A}} \times \bar{\mathcal{A}} \to \bar{\mathcal{A}}$. The corresponding ring homomorphism $W^{\mathrm{B}} \to W^{\mathrm{B}} \otimes W^{\mathrm{B}}$ is denoted by Δ_*. This picture can be filled into the $\mathcal{M} = \mathrm{Vec}_{\mathbb{Q}} \times_{\mathrm{Vec}_{\mathbb{C}}} \mathrm{Vec}_{\mathbb{Q}}$ world with a comparison isomorphism attached to an associator Φ. By the computation of the cohomology of algebroids we can show that the de Rham realization $\Delta_*\colon W^{\mathrm{DR}} \to W^{\mathrm{DR}} \otimes W^{\mathrm{DR}}$ is equal to the harmonic coproduct. By using the fake comparison isomorphism we have the theorem.

References

[De] Deligne, P., Le groupe fondamental de la droite projective moins trois points. In *Galois groups over Q* (Berkeley, CA, 1987), Math. Sci. Res. Inst. Publ. 16, Springer-Verlag, New York 1989, 79–297.

[DG] Deligne, P., Goncharov, A., Groupe fondamentaux motivique de Tate mixte. *Ann. Sci. Ecole Norm. Sup.* (4) **38** (2005), 1–56.

[DM] Deligne, P., Milne, J. S., Tannakian Categories. In *Hodge Cycles, Motives, and Shimura Varieties*, Lecture Notes in Math. 900, Springer-Verlag, Berlin 1982, 101–228.

[Dr] Drinfeld, V. G., On quasitriangular quasi-Hopf algebras and on a group that is closely connected with $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. *Leningrad Math. J.* **2** (1991), 829–860.

[G] Goncharov, A., Multiple polylogarithm and Mixed Tate motives. math.AG.0103059.

[H] Hoffman, M., Multiple harmonic series. *Pacific J. Math.* **152** (1992), 275–290.

[I] Ihara, Y., On the embedding of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ into $\widehat{\mathrm{GT}}$. In *Grothendieck theory of dessins d'enfants*, London Math. Soc. Lecture Note Ser. 200, Cambridge University Press, Cambridge 1994, 289–321.

[IKZ] Ihara, K., Kaneko, M., Zagier, D., Derivation and double shuffle relations for multiple zeta values. MPIM2004-100.

[L] Levine, M., *Mixed Motives*. Math. Surveys Monographs 57, Amer. Math. Soc., Providence, RI, 1998.

[R] Racinet, G., Doubles mélanges des polylogarithmes multiples aux racines de l'unité. *Publ. Math. Inst. Hautes Études Sci.* **95** (2002), 185–231.

[T] Terasoma, T., Mixed Tate motives and multiple zeta values. *Invent. Math.* **149** (2002), 339–369.

Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Tokyo, 153-8914 Japan

E-mail: terasoma@ms.u-tokyo.ac.jp

Geometry over nonclosed fields

Yuri Tschinkel

Abstract. I discuss some arithmetic aspects of higher-dimensional algebraic geometry. I focus on varieties with many rational points and on connections with classification theory and the minimal model program.

Mathematics Subject Classification (2000). Primary 14G05; Secondary 14G40.

Keywords. Rational and integral points, heights, rational curves, rational connectedness.

1. Introduction

Let k be a field, X an algebraic variety over k and $X(k)$ the set of k-rational points on X. Broadly speaking, arithmetic geometry is concerned with the set $X(k)$, but usually with additional structure, coming from the Zariski topology on X or the Galois group of k. Questions of arithmetic nature arise even in classical algebraic geometry over $\mathbb{C}$ – a rational section of a fibration $X \to B$ is a rational point of the generic fiber, considered as a variety over the function field $\mathbb{C}(B)$ of the base. Of particular importance are "small" ground fields k, such as finite fields $\mathbb{F}_p$ or the rational numbers $\mathbb{Q}$, as they are closest to the theory of diophantine equations, with its wealth of challenging problems.

The geometric approach to diophantine equations is inspired by the analogy "numbers – functions" going back at least to Riemann, Dedekind, Kronecker, Weber, Hensel, Weil, and many others. More concretely, one expects close relationships between global geometric properties of X, considered as an algebraic variety over "large" fields such as $\mathbb{C}$, and arithmetic properties. Sometimes one may need to stabilize the arithmetic situation by passing to finite or even infinite extensions of k, or by excluding exceptional subsets.

This area of arithmetic and geometry is dominated by the following themes:

- Existence of rational points. In dimension one this includes the work of Mazur and Merel on torsion points on elliptic curves, conjectures of Birch and Swinnerton-Dyer, Wiles' work on Fermat's conjecture, the theorems of Gross–Zagier, Kolyvagin and their generalizations. In higher dimensions, the attention is on the Brauer–Manin obstruction to the Hasse principle and weak approximation, its uniqueness and effective computation (see [54]).

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

- Density of rational points in various topologies, e.g., Zariski density or density in analytic topologies (see [51], [39]).
- Heights. This covers questions of bounding heights of points on higher genus curves (effective Mordell), finding lower bounds for heights of rational and algebraic points, Manin's conjecture about asymptotics of rational points of bounded height and its refinements (see [13], [55]).

The field is evolving rapidly, opening new directions of research and raising many concrete and interesting questions. In this survey I focus on varieties with many rational points and discuss connections with classification theory and the minimal model program in algebraic geometry.

Acknowledgments. I am deeply grateful to my teachers, friends and collaborators for sharing their ideas and insights, which over the years shaped my understanding of the subject: V. Batyrev, F. Bogomolov, A. Chambert-Loir, J. Franke, J. Harris, B. Hassett, J. Kollár, A. Kresch, Y. Manin, B. Mazur, E. Peyre, J. Shalika, M. Strauch, R. Takloo-Bighash.

2. Classification schemes

In this section we work over $\mathbb{C}$. Smooth projective algebraic varieties X can be classified by their topological and algebraic invariants: fundamental group $\pi_1(X)$, Brauer group $\mathrm{Br}(X)$, or a property of the canonical line bundle K_X:

$$\kappa(X) := \limsup_{n\to\infty} \log(h^0(X, nK_X))/\log(n) \in \{-\infty, 0, 1, \dots, \dim(X)\},$$

the Kodaira dimension. Alternatively, there are several geometric notions reflecting how close X is to a projective space $\mathbb{P}^n$: rationality, unirationality, uniruledness or rational connectedness, resp. chain-connectedness. Sometimes, X is realized by particularly simple equations, e.g., as a complete intersection in projective or weighted projective space. In this case, one could also classify by the degree $\deg(X)$ in this embedding. Sometimes, there is a distinguished polarization: (multiples of) the canonical, resp. the anticanonical line bundle; in the first case, $\kappa(X) = \dim(X)$, and one speaks of varieties of *general type*, and in the second, of *Fano* varieties. For historical reasons, Fano surfaces are called Del Pezzo surfaces. All other varieties are called varieties of *intermediate type*; an important subclass are varieties with trivial canonical class, e.g., abelian varieties and Calabi–Yau varieties.

The only Fano variety in dimension 1 is $\mathbb{P}^1$. Curves of general type are curves of genus ≥ 2. Smooth Del Pezzo surfaces divide into following types: $\mathbb{P}^2$, $\mathbb{P}^1 \times \mathbb{P}^1$, and blowups of $\mathbb{P}^2$ in ≤ 8 points in general position. In each dimension, there are only finitely many families of Fano varieties [49], [19].

The notions above are intertwined in many ways. For example, a smooth hypersurface $X \subset \mathbb{P}^n$ is Fano if $\deg(X) < n + 1$, and of general type if $\deg(X) > n + 1$,

so that we generally think of Fano varieties as being varieties of *small* degree, and geometrically not "too far" from $\mathbb{P}^n$. A Fano variety is rationally connected. If X is rationally connected, then it has trivial $\pi_1(X)$. If X is smooth and rational then $\mathrm{Br}(X)$ is trivial. A rationally connected X is rational, provided $\dim(X) \leq 2$, but not necessarily if $\dim(X) = 3$, etc. There are conjectured criteria linking algebraic and geometric properties, e.g., X is uniruled if and only if $\kappa(X) = -\infty$. A fundamental result is the following:

Theorem 2.1 (Graber–Harris–Starr [36]). *Let $\pi : X \to B$ be a morphism from a smooth projective variety to a smooth projective curve. Assume that the generic fiber is a rationally connected variety. Then π has a section.*

For applications, one needs to better understand the class of varieties of intermediate type. As a consequence of Theorem 2.1 and [49], [19], we know that every X admits an *MRC quotient*, a rational map $\pi_r : X \to Y = R(X)$ with rationally connected fibers and base not uniruled. If Y is not a point, this is followed by another fibration $Y \to Z$, with $\dim(Z) = \kappa(Y)$ and generic fiber of Kodaira dimension zero. However, Z is not necessarily of general type, and the construction may have to be iterated.

An alternative classification scheme has been proposed by Campana. A variety X is *special* if there are no (birational) dominant maps $\pi : X \to Y$, where Y is a smooth variety of $\dim(Y) \geq 1$ and of *orbi*-general type, with *orbi*-canonical class computed taking into account *multiplicities* of singular fibers of π. The class of special varieties includes rationally connected varieties and varieties of Kodaira dimension zero. Every X admits (birationally) a fibration $\pi_c : X \to Y = C(X)$ onto its *core* (or *le cœur*), with *special* generic fiber and base of orbi-general type [20]. For examples of simply connected X which do not admit birational maps onto varieties of general type of $\dim \geq 1$, but have a nontrivial core $C(X)$, see [10].

3. Potential density

One says that rational points on an algebraic variety X over a field k are *potentially dense* if there exists a finite extension k'/k such that $X(k')$ is Zariski dense in X. This birational property holds for projective spaces and abelian varieties and thus for varieties dominated by these, e.g., unirational varieties or Kummer varieties. It is preserved under étale covers of proper varieties. An important problem is to find necessary and sufficient geometric conditions for potential density of rational points. An extremal statement, almost an "axiom", is the following generalization of Mordell's conjecture to higher dimensions.

Conjecture 3.1 (Bombieri–Lang). Let X be a variety of general type over a number field k. Then $X(k)$ is contained in a proper subvariety, i.e., rational points on X are not potentially dense.

The outstanding result is Faltings' proof of this conjecture for subvarieties of abelian varieties [32], and in particular, curves of genus ≥ 2. Conjecture 3.1 has surprising consequences for *uniform* bounds for the number of rational points on curves of higher genus: it implies that there is a constant $c(k, \mathsf{g})$ such that for every curve C of genus $\mathsf{g} \geq 2$ over k one has $\#C(k) \leq c(k, \mathsf{g})$ [22].

In the opposite case of Fano varieties it is expected that their arithmetic properties are not too far from those of the projective space:

Conjecture 3.2 ([38]). Let X be a Fano variety over a number field k. Then rational points on X are potentially dense.

Conjecture 3.2 holds in dimension ≤ 2, for smooth quartic hypersurfaces of dimension ≥ 3 and for smooth hypersurfaces of degree 6 in the weighted projective space $\mathbb{P}(1, 1, 1, 2, 3)$ [38], [8]. There remains only one family of smooth Fano threefolds for which this question is open:

Problem 3.3. Let $X \to \mathbb{P}^3$ be a double cover ramified in a smooth surface S of degree 6. Show that rational points on X are potentially dense.

For singular S this has been resolved in [25].

The above conjectures are subsumed in

Conjecture 3.4 (Campana [20]). Let X be a smooth projective variety over a number field k. Rational points on X are potentially dense if and only if X is special.

An interesting subclass of special varieties to consider is the class of Calabi–Yau varieties, e.g., K3 surfaces and their higher-dimensional analogs – holomorphic symplectic varieties. Potential density holds for Enriques surfaces, elliptic K3 surfaces and K3 surfaces with infinite automorphism groups [9]; symmetric products of arbitrary K3 surfaces have been treated in [40]. Potential density holds *a posteriori* for varieties dominated by these.

Problem 3.5. Let X be a K3 surface over a number field, with geometric Picard number one. Show that rational points on X are potentially dense.

It would also be worthwhile to find nontrivial examples of three-dimensional Calabi–Yau varieties over number fields with dense sets of rational points.

The proofs of potential density rely on either automorphisms or on fibration structures, with fibers abelian varieties. It would be interesting to involve endomorphisms and to study orbits of the corresponding dynamical systems, in the spirit of [21]. Examples of quite nontrivial endomorphisms on certain holomorphic symplectic fourfolds are given in [67].

4. Points of bounded height

Assuming that rational points are Zariski dense one may seek some quantitative understanding of their distribution. A natural approach to this proceeds via the theory of heights. Let X be a projective algebraic variety over a number field k and $\mathcal{L} = (L, \|\cdot\|)$ an adelically metrized very ample line bundle on X (i.e., L is equipped with a family of v-adic norms, for each place v of k, which at almost all places are induced from a fixed global integral model, see [55]). Let

$$H_{\mathcal{L}}\colon X(k) \to \mathbb{R}_{>0}$$

be the associated height function, $X^\circ \subset X$ a subvariety of X and

$$\mathcal{N}(X^\circ, \mathcal{L}, B) := \#\{x \in X^\circ(k) \mid H_{\mathcal{L}}(x) \le B\} < \infty, \tag{1}$$

the counting function. The goal is to investigate its asymptotic behavior as $B \to \infty$, in terms of the pair $(X^\circ, \mathcal{L})$.

For example, let $X \subset \mathbb{P}^n$ be a hypersurface given by $f(\boldsymbol{x}) = 0$, where f is a homogeneous form in the variables $\boldsymbol{x} = (x_0, \dots, x_n)$ of degree d with coefficients in $\mathbb{Z}$. It is Fano for $n \ge d$. A counting function over $\mathbb{Q}$ is defined as

$$\mathcal{N}(X, \mathcal{O}(1), B) := \#\big\{\boldsymbol{x} \in (\mathbb{Z}^{n+1}_{\text{prim}} \setminus 0)/\pm,\ f(\boldsymbol{x}) = 0, \max_j |x_j| \le B\big\},$$

where $\mathbb{Z}^{n+1}_{\text{prim}}$ is the set of primitive vectors. The counting problem is to understand the asymptotic of this function as $B \to \infty$. The classical circle method solves this problem when the hypersurface X is smooth, the number of variables $n+1 \gg 2^d$ and if there exist solutions to $f(\boldsymbol{x}) = 0$ in all completions of $\mathbb{Q}$. Strong uniform bounds for points on hypersurfaces were recently established in [18], [46], [60]; these papers use in a crucial way the case of curves considered in [58], [14].

Problem 4.1. Let $X \subset \mathbb{P}^n$ be a smooth hypersurface of degree $d \le n$, over $\mathbb{Q}$. Show that there exists a Zariski open subset $X^\circ \subset X$ such that

$$\mathcal{N}(X^\circ, \mathcal{O}(1), B) = O(B^{n+1-d+\varepsilon}),$$

for all $\varepsilon > 0$. This is open already for $d = n = 3$.

In general, it is also difficult to produce rational points, even numerically. Interesting examples of varieties with *a priori* dense sets of rational points arise from group actions. More precisely, let G/k be a linear algebraic group, e.g., the Heisenberg group, the additive group $\mathbb{G}_a^d$ or the algebraic torus $\mathbb{G}_m^d$. Let

$$\rho\colon \mathsf{G} \to \mathrm{PGL}_{n+1} \tag{2}$$

be a representation over k. Fixing a point $\boldsymbol{x} \in \mathbb{P}^n(k)$, we can consider the *flow* $\rho(\mathsf{G})\cdot\boldsymbol{x}$ and count

$$\#\{\gamma \in \mathsf{G}(k) \mid H_{\mathcal{O}(1)}(\rho(\gamma)\cdot x) \le B\},$$

respectively, k-points in G/H, when the stabilizer H is nontrivial.

The asymptotic will depend on the group, the representation, the initial point $\boldsymbol{x}$, and the choice of the height. The difficulty is that already for $\mathsf{G} = \mathbb{G}_a^2$ there is no reasonable classification of possible representations ρ. This necessitates a change of language, from representation-theoretic to algebro-geometric: Let X be the Zariski closure of $\rho(\mathsf{G}) \cdot \boldsymbol{x} \subset \mathbb{P}^n$. Then X is

- an equivariant compactification of $X^\circ := \mathsf{G}/\mathsf{H}$,
- with a G-linearized very ample line bundle L,
- which is equipped with an adelic metrization $\mathscr{L} = (L, \|\cdot\|)$.

Then the counting problem is as in (1). Using equivariant resolution of singularities we can now reduce the counting problem to the case when X is smooth, and the boundary $X \setminus X^\circ$ is a divisor with normal crossings.

Over number fields, in available examples arising from group actions, such as flag varieties [33], toric varieties [4], horospherical varieties [66], equivariant compactification of $\mathbb{G}_a^n$ [23], De Concini–Procesi equivariant compactifications of semi-simple groups of adjoint type [64], [34], bi-equivariant compactifications of unipotent groups [63], the *algebraic-geometric* picture is as follows:

- the Picard group $\mathrm{Pic}(X)$ is a torsion free $\mathbb{Z}$-module,
- the (closed) cone of effective divisors $\Lambda_{\mathrm{eff}}(X) \subset \mathrm{Pic}(X)\otimes_{\mathbb{R}}$ is finitely generated,
- $-K_X$ is contained in the interior of $\Lambda_{\mathrm{eff}}(X)$.

The arithmetic picture, reflected in asymptotic formulas for the counting function (1), is described in terms of the geometric invariants as follows:

$$\mathscr{N}(X^\circ(k), \mathscr{L}, B) = c(\mathscr{L}) B^{a(L)} \log(B)^{b(L)-1}(1 + o(1)), \quad B \to \infty, \tag{3}$$

where

- $a(L) := \inf\{a \mid aL + K_X \in \Lambda_{\mathrm{eff}}(X)\}$,
- $b(L)$ is the maximal codimension of the face of $\Lambda_{\mathrm{eff}}(X)$ containing the class $a(L)L + K_X$.

In particular, $a(-K_X) = 1$ and $b(-K_X)$ is the rank of the Picard group of X, over k. The constant $c(\mathscr{L})$ depends on the choice of an adelic metrization, which determines the height. It was defined for $L = -K_X$ in [52] and in general in [5]:

$$c(-\mathcal{K}_X) := \alpha(X)\beta(X)\tau(\mathcal{K}_X) \tag{4}$$

where

- $\alpha(X)$ is the volume of certain polytope (intersection of the dual cone to $\Lambda_{\text{eff}}(X)$ with the affine hyperplane $\langle -K_X, \cdot \rangle = 1$),
- $\beta(X) = |\text{Br}(X)/\text{Br}(k)| = |H^1(\Gamma, \text{Pic}(\bar{X}))|$, is the order of the nontrivial part of the Brauer group,
- $\tau(-\mathcal{K}_X)$ is a Tamagawa type number.

For general polarizations, $c(\mathcal{L}) = \sum'_{y \in Y(k)} c(\mathcal{L}_y)$, where $X \to Y$ is a certain fibration arising in Fujita's version of the minimal model program; the summation is over a possibly infinite subset of the set of rational points on the base and $c(\mathcal{L}_y)$ are constants similar to (4).

Some of the above results have been extended to function fields $\mathbb{F}_q(B)$, where B is a curve over a finite field $\mathbb{F}_q$ [57], [16]. Over these fields, the constant $c(-\mathcal{K}_X)$ differs by an additional integral factor, $\neq 1$ already for some toric varieties [16].

Tamagawa numbers occurring here are natural generalizations of those in the theory of algebraic groups. An adelic metrization of K_X gives rise to v-adic measures ω_v on v-adic analytic manifolds $X(k_v)$. A regularized global measure $\omega_{\mathcal{K}}$ on the adeles $X(\mathbb{A}_k)$ is given by

$$\omega_{\mathcal{K}} := \text{L}^*_{\mathsf{S}}(1, \text{Pic}(X))|\text{disc}(k)|^{-\dim(X)/2} \prod_v \lambda_v^{-1} \omega_v, \tag{5}$$

where S is a finite set of valuations of k, including the archimedean ones, $\text{disc}(k)$ is the discriminant of k, the regularizing factors

$$\lambda_v := \begin{cases} \text{L}_v(1, \text{Pic}(X)) & v \notin \mathsf{S}, \\ 1 & v \in \mathsf{S}, \end{cases}$$

and

$$\text{L}^*_{\mathsf{S}}(1, \text{Pic}(X)) := \lim_{s \to 1} (s-1)^r \text{L}_{\mathsf{S}}(s, \text{Pic}(X))$$

is the leading coefficient at the pole of the partial Artin L-function associated to the Galois representation on the module $\text{Pic}(\bar{X})$. Then

$$\tau(-\mathcal{K}_X) := \int_{\overline{X(k)} \subset X(\mathbb{A}_k)} \omega_{\mathcal{K}}, \tag{6}$$

an integral over the closure of rational points in the adeles, in the direct product topology.

The proofs of these asymptotics use a combination of techniques from arithmetic geometry and analysis. A useful tool is the *height zeta function* defined by the series

$$\mathcal{Z}(X^\circ, \mathcal{L}, s) = \sum_{x \in X^\circ(k)} H_{\mathcal{L}}(x)^{-s}.$$

Tauberian theorems relate the asymptotic behavior of the counting function to analytic properties of the height zeta function. Analytic properties of $\mathcal{Z}$ are investigated via harmonic analysis on adelic groups or ergodic theory.

Problem 4.2. Prove Formula (3) for general equivariant compactifications of unipotent and solvable groups.

An alternative approach to asymptotics uses universal torsors. The essence of the method consists in the lifting of the counting problem for rational points on X to a counting problem for integral points on an auxiliary variety of higher dimension. This method has been used to reprove asymptotic results on toric varieties [60]. It has also been successfully applied to nonhomogeneous varieties, such as $\bar{\mathcal{M}}_{0,5}$ [17] and certain singular cubic surfaces [47], [26].

In the simplest case of a projective space, this can be explained by the familiar

$$\mathbb{Z}^{n+1}_{\mathrm{prim}} \setminus 0/\pm \xrightarrow{\mathbb{G}_m} \mathbb{P}^n(\mathbb{Q}),$$

certain integral points on the torsor $\mathbb{A}^{n+1} \setminus 0$, modulo units, are in bijection with rational points on the projective space. In general, the study of universal torsors leads to interesting algebraic problems, involving ideas from geometric invariant theory and toric geometry [2], [59], [42].

Problem 4.3. Compute equations of universal torsors for singular Del Pezzo surfaces in degrees $1, \dots, 4$.

The initial conjecture of Batyrev–Manin was that Formula (3) should hold for *all* Fano varieties, after enlarging the ground field, and restricting to appropriate Zariski open subsets X°. It is necessary to allow finite field extensions, since a Fano variety may not have any rational points over the ground field (e.g., a nonsplit conic). The restriction to Zariski open subsets is also necessary since the variety X may contain *accumulating* subvarieties, and the asymptotic of rational points on them can violate the conjecture (e.g, lines on a cubic surface). The Batyrev–Manin conjecture had to be adjusted, already in dimension 3, after the realization that certain fibrations may lead to differing $b(L)$ [3], [5]. The results mentioned above are convincing evidence, that the Batyrev–Manin conjecture and its refinement by Peyre should hold for Del Pezzo surfaces and for equivariant compactifications of all linear algebraic groups and their homogeneous spaces. A first step would be

Problem 4.4. Prove Formula (3) for singular Del Pezzo surfaces whose universal torsor is a hypersurface.

Problem 4.5. Prove Formula (3) for general spherical varieties over number fields.

5. Integral points

Let k be a number field, X a smooth projective algebraic variety over k and $D \subset X$ a divisor with strict normal crossings. In this log-geometric setup, it is the ampleness

of $(K_X + D)$, resp. $-(K_X + D)$, which characterizes the opposites in the classification picture. We need a theory of log-uniruledness, log-rational connectivity, and classification results and notions comparable to those in Section 2. e.g., log-special varieties etc. Like for rational points, one is interested in the distribution of integral points in Zariski topology and with respect to heights.

Choose models $\mathcal{X}$, $\mathcal{D}$ of X, D over the ring of integers $\mathfrak{o}_k$ of k. A rational point $x \in X(k)$ gives rise to a section $\pi_x \colon \mathrm{Spec}(\mathfrak{o}_k) \to \mathcal{X}$ of the structure morphism. Choose a finite set of places S of $\mathfrak{o}_k$. A $(\mathcal{D}, \mathsf{S})$-integral point is a section π_x such that for all $v \notin \mathsf{S}$ one has

$$\pi_{x,v} \cap \mathcal{D}_v = \emptyset,$$

i.e., π_x avoids $\mathcal{D}$ over $\mathrm{Spec}(\mathfrak{o}_k) \setminus \mathsf{S}$. We say that D-integral points on X are potentially dense, if there exists a finite extension k'/k, a finite set of places S' of $\mathfrak{o}_{k'}$, models $\mathcal{X}'$, $\mathcal{D}'$ of X, D over S' such that $(\mathcal{D}', \mathsf{S}')$-integral points in $X(k')$ are Zariski dense.

Conjecture 5.1 (Vojta). Assume that $K_X + D$ is ample. Then $(\mathcal{D}, \mathsf{S})$-integral points are contained in a proper subvariety, i.e., integral points are not potentially dense.

In analogy with Conjecture 3.2 we can formulate

Conjecture 5.2. Assume that $-(K_X + D)$ is ample. Then D-integral points on X are potentially dense.

An instance of what is expected in intermediate cases is the following "puncturing" conjecture, which could serve as a guiding principle:

Conjecture 5.3 ([41]). Let (X, Z) be a pair consisting of a smooth projective variety X and a smooth irreducible subvariety $Z \subset X$ of codimension ≥ 2, defined over a number field k. Let $\tilde{X} := \mathrm{Bl}_Z(X)$ be the blowup of X with center in Z, and D the exceptional divisor. Assume that rational points on X are potentially dense. Then D-integral points on $\tilde{X}$ are potentially dense.

Problem 5.4. Let X be an abelian variety of dimension ≥ 2 and Z a point. Prove potential density of D-integral points on the blowup $\tilde{X}$, where D is the exceptional divisor.

This holds if X is a product of at least two abelian varieties [41]. For numerical evidence in dimension 2, see [48].

Problem 5.5. Let X be a surface and $D \subset X$ a reduced effective Weil divisor such that the pair (X, D) has log-terminal singularities and $(K_X + D)$ is trivial. Show that D-integral points on X are potentially dense.

A special case, for $D = \emptyset$ and X smooth, is stated in Problem 3.5. Another open case arises for X a smooth Del Pezzo surface, and D a singular anticanonical curve.

When $X = \mathbb{P}^2$ this problem has been treated in [65] and [6]. Cubic surfaces X and smooth D have been considered in [7], this has been extended to smooth Del Pezzo surfaces X and smooth D in [41].

Assume now that $-(K_X + D)$ is ample and that $(\mathcal{D}, \mathsf{S})$-integral points are Zariski dense. Like in the case of rational points, group actions give a rich supply of quasi-projective projective varieties with many integral points. For example, replacing the projective representation ρ in (2) by a representation of an algebraic group G into GL_n and fixing $\mathbb{Z}$-structures leads to the study of points in $\mathsf{G}(\mathbb{Z})$, or integral points on the corresponding homogeneous spaces [27], [29], [30], [15].

For a Zariski open $X^\circ \subset X$, let

$$\mathcal{N}(X^\circ, -(\mathcal{K}_X + \mathcal{D}), B) := \#\{x \mid H_{-(\mathcal{K}_X+\mathcal{D})}(x) \le B\}$$

be the counting function on the set of $(\mathcal{D}, \mathsf{S})$-integral points in $X^\circ(k)$.

Our goal is a geometric interpretation of asymptotic formulas, similar to the one in Section 4 (see [24] for more details). Available asymptotics, proved via ergodic theory or harmonic analysis, are of the shape

$$\mathcal{N}(X^\circ, -(\mathcal{K}_X + \mathcal{D}), B) \sim c_{\mathsf{S}} B \log(B)^{b_S - 1}, \tag{7}$$

where

- $b_{\mathsf{S}} = \mathrm{rk}\ \mathrm{Pic}(X \setminus D) + \sum_{v \in \mathsf{S} \cup \mathsf{S}_\infty} r_v$;
- $r_v = \dim \mathrm{Cl}(\mathcal{D}_v)$, the dimension of the Clemens polytope associated to the set of boundary components of the reduction of $\mathcal{D}$ modulo v (vertices correspond to irreducible components, faces of codimension one to pairs of intersecting components, etc.);
- c_{S} is a constant similar to the one in (3), it involves a Tamagawa volume of the adeles outside S of $X \setminus D$ as in (6), with respect to the restriction of the Tamagawa measure from (5), and for each $v \in \mathsf{S} \cup \mathsf{S}_\infty$ and $\sigma \in \mathrm{Cl}_{\max}(\mathcal{D}_v)$, the Tamagawa volume of the closed subvariety $Z_{\sigma,v} \subset X(k_v)$, the intersection of corresponding components from σ. Here $\mathrm{Cl}_{\max}(\mathcal{D}_v)$ is the set of faces of the Clemens polytope of maximal dimension, and the Tamagawa measure on $Z_{\sigma,v}$ is obtained by adjunction.

Problem 5.6. Prove (7) for $X = \mathbb{P}^n$, $D \subset X$ a smooth hypersurface of low degree, defined by a form $f \in \mathbb{Z}[x_0, \dots, x_n]$, and S a finite set of primes of good reduction for D.

6. Arithmetic over function fields of curves

Here we work over $k = \mathbb{C}(B)$, where B is a smooth projective curve. Let X be a smooth projective variety over k and $\pi\colon \mathcal{X} \to B$ a model of X over B. Points in $X(k)$ correspond to sections of π, i.e., to certain curves in $\mathcal{X}$. A reformulation of Theorem 2.1 is

Theorem 6.1 ([36]). *Let X be a rationally connected variety over $k = \mathbb{C}(B)$. Then $X(k) \neq \emptyset$.*

One corollary of Theorem 2.1 is the proof of potential density, Conjecture 3.2, for rationally connected varieties over $k = \mathbb{C}(B)$. More precisely, after choosing finitely many smooth fibers and a point in each of these fibers one can find a section passing through these points [50]. This can be strengthened: after choosing finite *jets* (reductions of local analytic sections) in finitely many smooth fibers one can find a section which reduces to these jets [43]. It would be important to extend this property, called *weak approximation*, to singular fibers. Careful analysis of desingularizations of compound du Val singularities should yield a solution to

Problem 6.2. Prove weak approximation for cubic surfaces over function fields of curves.

Another step beyond what can be currently proved over number fields, are examples of K3 surfaces, and more generally, Calabi–Yau varieties, of geometric Picard number one with Zariski dense sets of rational points in [45].

Problem 6.3. Prove potential density of rational points for all K3 surfaces over $k = \mathbb{C}(B)$.

There are analogous questions for log-Fano varieties. For example,

Problem 6.4. Prove potential density of integral points on log-Fano varieties (Conjecture 5.2) over function fields of curves.

One of the main advantages of the field $k = \mathbb{C}(B)$ is that curves often deform. This opens the door for a systematic application of deformation theory, the theory of moduli spaces of stable curves and maps etc. The proofs of properties like potential density or weak approximation proceed by finding and smoothing special chains of rational curves (*combs* and *combs with broken teeth*).

An alternative approach to potential density, based on endomorphisms from [67], leads to examples of holomorphic symplectic fourfolds X over $\mathbb{C}(x)$ with geometric Picard number one and dense rational points [1].

7. Geometry over finite fields

In some aspects, finite fields are similar to function fields of curves. For example, an analog of Theorem 6.1 is:

Theorem 7.1 ([31]). *Let X be a smooth projective rationally connected variety over a finite field $k = \mathbb{F}_q$. Then $X(k) \neq \emptyset$.*

On the other hand, the classification schemes as outlined in Section 2 have to be adjusted. In particular, the relation between the ampleness of the canonical, resp. anticanonical, line bundle and rational connectedness is much less clear. For example, in positive characteristic, there exist unirational, and thus rationally connected, varieties of general type. But over finite fields $k = \mathbb{F}_q$, and their closures $\bar{k}$, there are also examples of surfaces X of general type which are not uniruled, have a nontrivial Brauer group, nontrivial fundamental group, and still have the property that on some dense Zariski open subset $X^\circ \subset X$, every finite set of $\bar{k}$-points lies a geometrically irreducible rational curve in X, defined over k. Moreover, every Kummer K3 surface over a finite field has this rather strong rational connectedness property [11], [12].

Problem 7.2. Let X be an elliptic K3 surface over a sufficiently large finite field. Show that X is rationally connected (in the above sense).

References

[1] Amerik, E., Campana, F., Fibrations meromorphes sur certaines variétés de classe canonique triviale. math.AG/0510299, 2005.

[2] Batyrev, V., Popov, O., The Cox ring of a Del Pezzo surface. In *Arithmetic of higher-dimensional algebraic varieties*, Progr. Math. 226, Birkhäuser, Boston, MA, 2004, 85–103.

[3] Batyrev, V., Tschinkel, Y., Rational points on some Fano cubic bundles. *C. R. Acad. Sci. Paris Sér. I Math.* **323** (1996), 41–46.

[4] Batyrev, V., Tschinkel, Y., Manin's conjecture for toric varieties. *J. Algebraic Geom.* **7** (1) (1998), 15–53.

[5] Batyrev, V., Tschinkel, Y., Tamagawa numbers of polarized algebraic varieties. *Astérisque* **251** (1998), 299–340.

[6] Beukers, F., Ternary form equations. *J. Number Theory* **54** (1) (1995), 113–133.

[7] Beukers, F., Integral points on cubic surfaces. In *Number theory* (Ottawa, ON, 1996), CRM Proc. Lecture Notes 19, Amer. Math. Soc., Providence, RI, 1999, 25–33.

[8] Bogomolov, F., Tschinkel, Y., On the density of rational points on elliptic fibrations. *J. Reine Angew. Math.* **511** (1999), 87–93.

[9] Bogomolov, F., Tschinkel, Y., Density of rational points on elliptic K3 surfaces. *Asian J. Math.* **4** (2) (2000), 351–368.

[10] Bogomolov, F., Tschinkel, Y., Special elliptic fibrations. In *The Fano Conference* (Torino, 2002), Università di Torino, Turin 2004, 223–234.

[11] Bogomolov, F., Tschinkel, Y., Rational curves and points on K3 surfaces. *Amer. J. Math.* **127** (4) (2005), 825–835.

[12] Bogomolov, F., Tschinkel, Y., Curves in abelian varieties over finite fields. *Internat. Math. Res. Notices* **4** (2005), 233–238.

[13] Bombieri, E., Gubler, W., *Heights in Diophantine Geometry*. New Math. Monogr. 4, Cambridge University Press, Cambridge 2006.

[14] Bombieri, E., Pila, J., The number of integral points on arcs and ovals. *Duke Math. J.* **59** (1989), 337–357.

[15] Borovoi, M., Rudnick, Z., Hardy-Littlewood varieties and semisimple groups. *Invent. Math.* **119** (1) (1995), 37–66.

[16] Bourqui, D., Constante de Peyre des variétés toriques en caractéristique positive. math.NT/0501409, 2005.

[17] de la Bretèche, R., Nombre de points de hauteur bornée sur les surfaces de Del Pezzo de degré 5. *Duke Math. J.* **113** (3) (2002), 421–464.

[18] Browning, T., Heath-Brown, R., Counting rational points on hypersurfaces. *J. Reine Angew. Math.* **584** (2005), 83–115.

[19] Campana, F., Connexité rationnelle des variétés de Fano. *Ann. Sci. École Norm. Sup.* (4) **25** (5) (1992), 539–545.

[20] Campana, F., Orbifolds, special varieties and classification theory. *Ann. Inst. Fourier* **54** (3) (2004), 499–630.

[21] Cantat, S., Sur la dynamique du groupe d'automorphismes des surfaces $K3$. *Transform. Groups* **6** (3) (2001), 201–214.

[22] Caporaso, L., Harris, J., Mazur, B., Uniformity of rational points. *J. Amer. Math. Soc.* **10** (1) (1997), 1–35.

[23] Chambert-Loir, A., Tschinkel, Y., On the distribution of points of bounded height on equivariant compactifications of vector groups. *Invent. Math.* **148** (2) (2002), 421–452.

[24] Chambert-Loir, A., Tschinkel, Y., On the distribution of integral points of bounded height. In preparation.

[25] Cheltsov, I., Park, J., Sextic double solids. math.AG/0404452, 2004.

[26] Derenthal, U., Manin's conjecture for a certain singular cubic surface. math.NT/0504016, 2005.

[27] Duke, W., Rudnick, Z., Sarnak, P., Density of integer points on affine homogeneous varieties. *Duke Math. J.* **71** (1) (1993), 143–179.

[28] Eskin, A., Counting problems and semisimple groups. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, Extra Vol. II, 539–552.

[29] Eskin, A., McMullen, C., Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.* **71** (1) (1993), 181–209.

[30] Eskin, A., Mozes, S., Shah, N., Unipotent flows and counting lattice points on homogeneous varieties. *Ann. of Math.* (2) **143** (2) (1996), 253–299.

[31] Esnault, H., Varieties over a finite field with trivial Chow group of 0-cycles have a rational point. *Invent. Math.* **151** (1) (2003), 187–191.

[32] Faltings, G., Diophantine approximation on abelian varieties. *Ann. of Math.* (2) **133** (3) (1991), 549–576.

[33] Franke, J., Manin, Y.I., Tschinkel, Y., Rational points of bounded height on Fano varieties. *Invent. Math.* **95** (1989), 421–435.

[34] Gorodnik, A., Maucourant, F., Oh, H., Manin's conjecture on rational points of bounded height and adelic mixing. Preprint, 2005.

[35] Graber, T., Harris, J., Mazur, B., Starr, J., Arithmetic questions related to rationally connected varieties. In *The legacy of Niels Henrik Abel*, Springer-Verlag, Berlin 2004, 531–542.

[36] Graber, T., Harris, J., Starr, J., Families of rationally connected varieties. *J. Amer. Math. Soc.* **16** (2003), 57–67.

[37] Graber, T., Harris, J., Starr, J., Rational connectivity and sections of families over curves. *Ann. Sci. École Norm. Sup.*, to appear.

[38] Harris, J., Tschinkel, Y., Rational points on quartics. *Duke Math. J.* **104** (3) (2000), 477–500.

[39] Hassett, B., Potential density of rational points on algebraic varieties. In *Higher dimensional varieties and rational points* (Budapest, 2001), Bolyai Soc. Math. Stud. 12, Springer-Verlag, Berlin 2003, 223–282.

[40] Hassett, B., Tschinkel, Y., Abelian fibrations and rational points on symmetric products. *Internat. J. Math.* **11** (9) (2000), 1163–1176.

[41] Hassett, B., Tschinkel, Y., Density of integral points on algebraic varieties. In *Rational points on algebraic varieties*, Progr. Math. 199, Birkhäuser, Basel 2001, 169–197.

[42] Hassett, B., Tschinkel, Y., Universal torsors and Cox rings. In *Arithmetic of higher-dimensional algebraic varieties* (Palo Alto, CA, 2002), Progr. Math. 226, Birkhäuser, Boston, MA, 2004, 149–173.

[43] Hassett, B., Tschinkel, Y., Weak approximation over function fields. *Invent. Math.* **163** (1) (2006), 171–190.

[44] Hassett, B., Tschinkel, Y., Approximation at places of bad reduction for rationally connected varieties. Preprint, 2005.

[45] Hassett, B., Tschinkel, Y., Potential density of rational points for K3 surfaces over function fields. Preprint, 2005.

[46] Heath-Brown, D. R., The density of rational points on curves and surfaces. *Ann. of Math.* (2) **155** (2) (2002), 553–595.

[47] Heath-Brown, D. R., The density of rational points on Cayley's cubic surface. In *Proceedings of the Session in Analytic Number Theory and Diophantine Equations*, Bonner Math. Schriften 360, Univ. Bonn, Bonn 2003.

[48] Kresch, A., Tschinkel, Y., Integral points on punctured abelian surfaces. In *Algorithmic number theory* (Sydney, 2002), Lecture Notes in Comput. Sci. 2369, Springer-Verlag, Berlin 2002, 198–204.

[49] Kollár, J., Miyaoka, Y., Mori, S., Rational connectedness and boundedness of Fano manifolds. *J. Differential Geom.* **36** (3) (1992), 765–779.

[50] Kollár, J., Miyaoka, Y., Mori, S., Rationally connected varieties. *J. Algebraic Geom.*, **1** (3) (1992), 429–448.

[51] Mazur, B., Open problems regarding rational points on curves and varieties. In *Galois representations in arithmetic algebraic geometry* (Durham, 1996), London Math. Soc. Lecture Note Ser. 254, Cambridge University Press, Cambridge 1998, 239–265.

[52] Peyre, E., Hauteurs et nombres de Tamagawa sur les variétés de Fano. *Duke Math. J.* **79** (1995), 101–128.

[53] Peyre, E., Terme principal de la fonction zeta des hauteurs et torseurs universels. *Astérisque* **251** (1998), 259–298.

[54] Peyre, E., Counting points on varieties using universal torsors. In *Arithmetic of higher-dimensional algebraic varieties* (Palo Alto, CA, 2002), Progr. Math. 226, Birkhäuser, Boston, MA, 2004, 61–81.

[55] Peyre, E., Points de hauteur bornée et géométrie des variétés (d'après Y. Manin et al.). In *Séminaire Bourbaki*, Vol. 2000/2001; *Astérisque* **282** (2002), Exp. No. 891, ix, 323–344.

[56] Peyre, E., Obstructions au principe de Hasse et à l'approximation faible. *Séminaire Bourbaki*, Vol. 2003/2004; *Astérisque* **299** (2005), Exp. No. 931, viii, 165–193.

[57] Peyre, E., Points de hauteur bornee sur les varietes de drapeaux en caracteristique finie. math.NT/0303067, 2003.

[58] Pila, J., Density of integral and rational points on varieties. *Astérisque* **228** (1995), 183–187.

[59] Popov, O., The Cox ring of a Del Pezzo surface has rational singularities. math.AG/0402154, 2004.

[60] Salberger, P., Tamagawa measures on universal torsors and points of bounded height on Fano varieties. *Astérisque* **251** (1998), 91–258.

[61] Salberger, P., Counting rational points on hypersurfaces of low dimension. *Ann. Sci. École Norm. Sup.* (4) **38** (1) (2005), 93–115.

[62] Shalika, J., Tschinkel, Y., Height zeta functions of equivariant compactifications of the Heisenberg group. In *Contributions to Automorphic Forms, Geometry, and Number Theory*, Johns Hopkins University Press, Baltimore, MD, 2004, 743–771.

[63] Shalika, J., Tschinkel, Y., Height zeta functions of equivariant compactifications of unipotent groups. In preparation.

[64] Shalika, J., Takloo-Bighash, R., Tschinkel, Y., Rational points on compactifications of semi-simple groups. Preprint, 2005.

[65] Silverman, J., Integral points on curves and surfaces. *Number theory* (Ulm, 1987), 202–241, Lecture Notes in Math. 1380, Springer-Verlag, Berlin 1989.

[66] Strauch, M., Tschinkel, Y., Height zeta functions of toric bundles over flag varieties. *Selecta Math.* (*N.S.*) **5** (3) (1999), 325–396.

[67] Voisin, C., Intrinsic pseudo-volume forms and K-correspondences. In *The Fano Conference* (Torino, 2002), Università di Torino, Turin 2004, 761–792.

Courant Institute of Mathematical Sciences, 251 Mercer St., New York, NY 10012, U.S.A.
and
Mathematisches Institut, Bunsenstr. 3-5, 37073 Göttingen, Germany
E-mail: tschinkel@cims.nyu.edu

Algebraic Morse theory and the weak factorization theorem

Jarosław Włodarczyk*

Abstract. We develop a Morse-like theory for complex algebraic varieties. In this theory a Morse function is replaced by a $\mathbb{C}^*$-action. The critical points of the Morse function correspond to connected fixed point components. "Passing through the fixed points" induces some simple birational transformations called blow-ups, blow-downs and flips which are analogous to spherical modifications.

In classical Morse theory by means of a Morse function we can decompose the manifold into elementary pieces – "handles". In algebraic Morse theory we decompose a birational map between two smooth complex algebraic varieties into a sequence of blow-ups and blow-downs with smooth centers.

Mathematics Subject Classification (2000). Primary 14E05.

Keywords. Birational maps, blow-ups, $\mathbb{C}^*$-actions, toric varieties.

1. Introduction

We shall work over an algebraically closed field K of characteristic zero. We denote by K^* the multiplicative group of K.

In this paper we outline our proof of the following theorem:

Theorem 1.1 (The Weak Factorization Theorem). 1. *Let $f : X \dashrightarrow Y$ be a birational map of smooth complete varieties over a field of characteristic zero, which is an isomorphism over an open set U. Then f can be factored as*

$$X = X_0 \overset{f_0}{\dashrightarrow} X_1 \overset{f_1}{\dashrightarrow} \cdots \overset{f_{n-1}}{\dashrightarrow} X_n = Y,$$

where each X_i is a smooth complete variety and f_i is a blow-up or blow-down at a smooth center which is an isomorphism over U.

2. *Moreover, if $X \setminus U$ and $Y \setminus U$ are divisors with simple normal crossings, then each $D_i := X_i \setminus U$ is a divisor with simple normal crossings and f_i is a blow-up or blow-down at a smooth center which has normal crossings with components of D_i.*

3. *There is an index $1 \leq r \leq n$ such that for all $i \leq r$ the induced birational map $X_i \overset{f_0}{\dashrightarrow} X$ is a projective morphism and for all $r \leq i \leq n$ the map $X_i \overset{f_0}{\dashrightarrow} Y$ is projective morphism.*

*The author was supported in part by NSF grant DMS-0500659 and Polish KBN grant GR-1784.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

4. *The above factorization commutes with any automorphisms ϕ_X of X, and ϕ_Y of Y such that $f \circ \phi_X = \phi_Y \circ f$.*

The theorem was proven in [38] and in [3] in a more general version. The above formulation essentially reflects the statement of the theorem in [3].

The weak factorization theorem extends a theorem of Zariski, which states that any birational map between two smooth complete surfaces can be factored into a succession of blow-ups at points followed by a succession of blow-downs at points. A stronger version of the above theorem, called the strong factorization conjecture, remains open.

Conjecture 1.2 (Strong Factorization Conjecture). Any birational map $f\colon X \dashrightarrow Y$ of smooth complete varieties can be factored into a succession of blow-ups at smooth centers followed by a succession of blow-downs at smooth centers.

Note that both statements are equivalent in dimension 2. One can find the formulation of the relevant conjectures in many papers. Hironaka [17] formulated the strong factorization conjecture. The weak factorization problem was stated by Miyake and Oda [30]. The toric versions of the strong and weak factorizations were also conjectured by Miyake and Oda [30] and are called the strong and weak Oda conjectures. The 3-dimensional toric version of the weak form was established by Danilov [12] (see also Ewald [14]). The weak toric conjecture in arbitrary dimensions was proved in [36] and later independently by Morelli [27], who also claimed to have a proof of the strong factorization conjecture (see also Morelli [28]). Morelli's proof of the weak Oda conjecture was completed, revised and generalized to the toroidal case by Abramovich, Matsuki and Rashid in [4]. A gap in Morelli's proof of the strong Oda conjecture, which went unnoticed in [4], was later found by K. Karu.

The local version of the strong factorization problem was posed by Abhyankar in dimension 2 and by Christensen in general; Christensen has solved it for 3-dimensional toric varieties [8]. The local version of the weak factorization problem (in characteristic 0) was solved by Cutkosky [9], who also showed that Oda's strong conjecture implies the local version of the strong conjecture for proper birational morphisms [10] and proved the local strong factorization conjecture in dimension 3 ([10]) via Christensen's theorem. Finally Karu generalized Christensen's result to any dimension and completed the argument for the local strong factorization ([22]).

The proofs in [38] and [3] are both build upon the idea of cobordism which was developed in [37] and was inspired by Morelli's theory of polyhedral cobordisms [27]. The main idea of [37] is to construct a space with a K^*-action for a given birational map. The space called a birational cobordism resembles the idea of Morse cobordism and determines a decomposition of the birational map into elementary transformations (see Remark 2.6). This gives a factorization into a sequence of weighted blow-ups and blow-downs. One can view the birational maps determined by cobordisms also in terms of VGIT developed in papers of Thaddeus ([34]) and Dolgachev–Hu ([13]). As shown in [37] the weighted blow-ups which occur in the factorization have

locally a natural toric and combinatorial description which is crucial for their further regularization.

The two existing methods of regularizing centers of this factorization are π-desingularization of cobordisms as in [38] and local torification of the action as in [3].

The present proof is essentially the same as in [38]. Instead of working in full generality and developing the suitable language for toroidal varieties we focus on applying the general ideas to a particular construction of a smooth cobordism. The reader can find also a more general and extended version of this proof in [39]. The π-desingularization is a desingularization of geometric quotients of a K^*-action. This can be done locally and the procedure can be globalized in the functorial and even canonical way. The π-desingularization makes all the intermediate varieties (which are geometric quotients) smooth, and also the connecting blow-ups have smooth centers.

The proof of Abramovich, Karu, Matsuki and the author [3] relies on a subtle analysis of differences between locally toric and toroidal structures defined by the action of K^*. The Abramovich–de Jong idea of torification is roughly speaking to construct the ideal sheaves whose blow-ups (or principalizations) introduce the structure of toroidal varieties in neighborhoods of fixed points of the action. This allows one to pass from birational maps between intermediate varieties in the neighborhood of fixed points to birational toroidal maps. The latter can be factored into a sequence of smooth blow-ups by using the same combinatorial methods as for toric varieties. Combining all the local factorizations together we get a global factorization.

For simplicity we restrict our considerations to projective varieties. We will not discuss here compatibility with divisors and functorial properties.

2. Birational cobordisms

2.1. Definition of a birational cobordism. Recall some basic definitions from Mumford's GIT theory.

Definition 2.1. Let K^* act on X. By a *good quotient* we mean a variety $Y = X//K^*$ together with a morphism $\pi\colon X \to Y$ which is constant on G-orbits such that for any affine open subset $U \subset Y$ the inverse image $\pi^{-1}(U)$ is affine and $\pi^*\colon O_Y(U) \to O_X(\pi^{-1}(U))^{K^*}$ is an isomorphism. If additionally for any closed point $y \in Y$ its inverse limit $\pi^{-1}(x)$ is a single orbit we call $Y := X/K^*$ together with $\pi\colon X \to Y$ a *geometric quotient*.

Remark 2.2. A geometric quotient is a space of orbits while a good quotient is a space of equivalence classes of orbits generated by the relation that two orbits are equivalent if their closures intersect.

Definition 2.3. Let K^* act on X. We say that $\lim_{t\to 0} tx$ exists (respectively $\lim_{t\to\infty} tx$ exists) if the morphism $K^* \to X$ given by $t \mapsto tx$ extends to $K^* \subset \mathbb{A}^1 \to X$ (or respectively $K^* \subset \mathbb{P}^1 \setminus \{0\} \to X$).

Definition 2.4 ([37]). Let X_1 and X_2 be two birationally equivalent normal varieties. A *birational cobordism* or simply a *cobordism* $B := B(X_1, X_2)$ between them is a normal variety B with an algebraic action of K^* such that the sets

$$B_- := \{x \in B \mid \lim_{t\to 0} tx \text{ does not exist}$$

and

$$B_+ := \{x \in B \mid \lim_{t\to \infty} tx \text{ does not exist}\}$$

are nonempty and open and there exist geometric quotients B_-/K^* and B_+/K^* such that $B_+/K^* \simeq X_1$ and $B_-/K^* \simeq X_2$ and the birational map $X_1 \dashrightarrow X_2$ is given by the above isomorphisms and the open embeddings of $B_+ \cap B_-/K^*$ into B_+/K^* and B_-/K^* respectively.

Remark 2.5. An analogous notion of cobordism of fans of toric varieties was introduced by Morelli in [27].

Remark 2.6. The above definition can also be considered as an analog of the notion of cobordism in Morse theory. Let W be a cobordism in Morse theory of two differentiable manifolds X and X' and $f\colon W \to [a, b] \subset \mathbb{R}$ be a Morse function such that $f^{-1}(a) = X$ and $f^{-1}(b) = X'$. Then X and X' have open neighborhoods $X \subseteq V \subseteq W$ and $X' \subseteq V' \subseteq W'$ such that $V \simeq X \times [a, a + \varepsilon)$ and $V' \simeq X' \times (b - \varepsilon, b]$ for which $f_{|V}\colon V \simeq X \times [a, a + \varepsilon) \to [a, b]$ and $f_{|V'}\colon V' \simeq X' \times (b - \varepsilon, b] \to [a, b]$ are the natural projections on the second coordinate. Let $W' := W \cup_V X \times (-\infty, a + \varepsilon) \cup_{V'} X' \times (b - \varepsilon, +\infty)$. One can easily see that W' is isomorphic to $W \setminus X \setminus X' = \{x \in W \mid a < f(x) < b\}$. Let $f'\colon W' \to \mathbb{R}$ be the map defined by glueing the function f and the natural projection on the second coordinate. Then $\operatorname{grad}(f')$ defines an action on W' of a 1-parameter group $T \simeq \mathbb{R} \simeq \mathbb{R}^*_{>0}$ of diffeomorphisms. The last group isomorphism is given by the exponential.

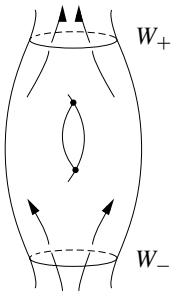


Figure 1. Cobordism in Morse theory.

Then one can see that $W'_- := \{x \in W' \mid \lim_{t\to 0} tx \text{ does not exist}\}$ and $W'_+ := \{x \in W' \mid \lim_{t\to\infty} tx \text{ does not exist}\}$ are open and X and X' can be considered as

quotients of these sets by T. The critical points of the Morse function are T-fixed points. "Passing through the fixed points" of the action induces a simple birational transformation similar to spherical modification in Morse theory (see Example 2.7).

Example 2.7. Let K^* act on $B := \mathbb{A}_K^{l+m+r}$ by

$$t(x_1, \dots, x_l, y_1, \dots, y_m, z_1, \dots, z_r) \\ = (t^{a_1} \cdot x_1, \dots, t^{a_l} \cdot x_l, t^{-b_1} \cdot y_1, \dots, t^{-b_m} \cdot y_m, z_1, \dots, z_r),$$

where $a_1, \dots, a_l, b_1, \dots, b_m > 0$. Set $\bar{x} = (x_1, \dots, x_l)$, $\bar{y} = (y_1, \dots, y_m)$, $\bar{z} = (z_1, \dots, z_r)$. Then

$$B_- = \{p = (\bar{x}, \bar{y}, \bar{z}) \in \mathbb{A}_K^{l+m+r} \mid \bar{y} \neq 0\},$$
$$B_+ = \{p = (\bar{x}, \bar{y}, \bar{z}) \in \mathbb{A}_K^{l+m+r} \mid \bar{x} \neq 0\}.$$

Case 1. $a_i = b_i = 1$, $r = 0$ (Atiyah, Reid). One can easily see that $B//K^*$ is the affine cone over the Segre embedding $\mathbb{P}^{l-1} \times \mathbb{P}^{m-1} \to \mathbb{P}^{l \cdot m - 1}$, and B_+/K^* and B_-/K^* are smooth.

The relevant birational map $\phi\colon B_-/K^* \dashrightarrow B_+/K^*$ is a flip for $l, m \geq 2$ replacing $\mathbb{P}^{l-1} \subset B_-/K^*$ with $\mathbb{P}^{m-1} \subset B_+/K^*$. For $l = 1$, $m \geq 2$, ϕ is a blow-down, and for $l \geq 2$, $m = 1$ it is a blow-up. If $l = m = 1$ then ϕ is the identity. One can show that $\phi\colon B_-/K^* \dashrightarrow B_+/K^*$ factors into the blow-up of $\mathbb{P}^{l-1} \subset B_-/K^*$ followed by the blow-down of $\mathbb{P}^{m-1} \subset B_+/K^*$.

Case 2. General case. For $l = 1$, $m \geq 2$, ϕ is a toric blow-up whose exceptional fibers are weighted projective spaces. For $l \geq 2$, $m = 1$, ϕ is a toric blow-down. If $l = m = 1$ then ϕ is the identity. The birational map $\phi\colon B_-/K^* \dashrightarrow B_+/K^*$ factors into a weighted blow-up and a weighted blow-down.

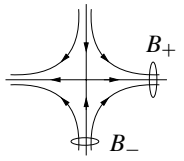


Figure 2. Affine Cobordism.

Remark 2.8. In Morse theory we have an analogous situation. In cobordisms with one critical point we replace S^{l-1} by S^{m-1}.

2.2. Fixed points of the action. Let X be a variety with an action of K^*. Denote by X^{K^*} the set of fixed points of the action and by $\mathcal{C}(X^{K^*})$ the set of its irreducible fixed components. For any $F \in \mathcal{C}(X^{K^*})$ set

$$F^+(X) = F^+ = \{x \in X \mid \lim_{t \to 0} tx \in F\}, \quad F^-(X) = F^- = \{x \in X \mid \lim_{t \to \infty} tx \in F\}.$$

Example 2.9. In Example 2.7,

$$F = \{p \in B \mid \bar{x} = \bar{y} = 0\}, \quad F^- = \{p \in B \mid \bar{x} = 0\}, \quad F^+ = \{p \in B \mid \bar{y} = 0\}.$$

Lemma 2.10. *If F is the fixed point set of an affine variety U then F, F^+ and F^- are closed in U. Moreover the ideals $I_{F^+}, I_{F^-} \subset K[V]$ are generated by all semiinvariant functions with positive (respectively negative) weights.*

Proof. Embed U equivariantly into affine space $\mathbb{A}^n$ with linear action and use the example above. □

2.3. Existence of a smooth birational cobordism

Proposition 2.11. *Let $\phi\colon X \dashrightarrow Y$ be a birational map between smooth projective varieties. Then ϕ factors as $X \leftarrow Z \to Y$, where $Z \to X$ and $Z \to Y$ are birational morphisms from a smooth projective variety Z.*

Proof. Let $\Gamma(X, Y) \subset X \times Y$ be the graph of ϕ and Z be its canonical resolution of singularities [17]. □

It suffices to construct the cobordism and factorization for the projective morphism $Y \to X$.

Proposition 2.12 ([37]). *Let $\varphi\colon Y \to X$ be a birational morphism of smooth projective varieties with the exceptional divisor D. Let $U \subset X, Y$ be an open subset where φ is an isomorphism. There exists a smooth projective variety $\overline{B}$ with a K^*-action, which contains fixed point components isomorphic to X and Y such that*

- *$B = B(X, Y) := \overline{B} \setminus (X \cup Y)$ is a cobordism between X and Y;*
- *$U \times K^* \subset B_- \cap B_+ \subset B$;*
- *there are K^*-equivariant isomorphisms $X^- \simeq X \times (\mathbb{P}^1 \setminus \{0\})$ and $Y^+ \simeq \mathcal{O}_Y(D)$;*
- *$X^- \setminus X = B_+$ and $Y^+ \setminus Y = B_-$*

In further considerations we shall refer to $\overline{B}$ as a *compactified cobordism.*

Proof. We follow here the Abramovich construction of cobordism. Let $\mathcal{I} \subset \mathcal{O}_X$ be a sheaf of ideals such that $Y = \mathrm{Bl}_{\mathcal{I}} X$ is obtained from X by blowing up of $\mathcal{I}$. Let z denote the standard coordinate on $\mathbb{P}^1$ and let $\mathcal{I}_0$ be the ideal of the point $z = 0$ on $\mathbb{P}^1$. Set $W := X \times \mathbb{P}^1$ and denote by $\pi_1\colon W \to X, \pi_2\colon W \to \mathbb{P}^1$ the standard projections. Then $\mathcal{J} := \pi_1^*(\mathcal{I}) + \pi_2^*(\mathcal{I}_0)$ is an ideal supported on $X \times \{0\}$. Set $W' := \mathrm{Bl}_{\mathcal{J}} W$. The proper transform of $X \times \{0\}$ is isomorphic to Y and we identify it with Y. Let us describe Y locally. Let $f_1, \ldots, f_k$ generate the ideal $\mathcal{I}$ on some open affine set $U \subset X$. Then after the blow-up $Y \to X$ at $\mathcal{I}$ the inverse image of U is a union of open charts $U_i \subset Y$, where

$$K[U_i] = K[U][f_1/f_i, \ldots, f_k/f_i].$$

Now the functions $f_1, \ldots, f_k, z$ generate the ideal $\mathcal{J}$ on $U \times \mathbb{A}^1 \subset W$. After the blow-up $W' \to W$ at $\mathcal{J}$, the inverse image of $U \times \mathbb{A}^1$ is a union of open charts $V_i \supset Y$, where

$$K[V_i] = K[U][f_1/f_i, \ldots, f_k/f_i, z/f_i] = K[U_i][z/f_i]$$

and the relevant V_z which does not intersect Y. Then $V_i = U_i^+ \simeq U_i \times \mathbb{A}^1$ where $z' := z/f_i$ is the standard coordinate on $\mathbb{A}^1$. The action of K^* on the factor U is trivial while on $\mathbb{A}^1$ it is standard given by $t(z') = tz$. Thus the open subset $Y^+ = \bigcup U_i^+ = \bigcup V_i \subset W'$ is a line bundle over Y with the standard action of K^*. On the other hand the neighborhood $X^- := X \times (\mathbb{P}^1 \setminus \{0\})$ of $X \subset W$ remains unchanged after the blow-up of $\mathcal{J}$. We identify X with $X \times \{\infty\}$. We define $\bar{B}$ to be the canonical desingularization of W. Then $B := \bar{B} \setminus X \setminus Y$. We get $B_-/K^* = (Y^+ \setminus Y)/K^* = Y$, while $B_+/K^* = (X^+ \setminus X)/K^* = X$. □

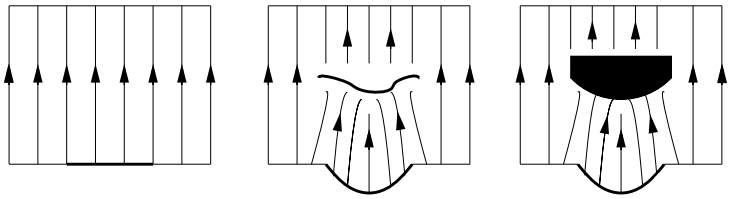


Figure 3. Compactified cobordism.

Remark 2.13. The Abramovich construction can be considered as a generalization of the Fulton–Macpherson construction of the deformation to the normal cone. If we let $\mathcal{I} = \mathcal{I}_C$ be the ideal sheaf of the smooth center then the relevant blow-up is already smooth. On the other hand this is a particular case of the very first example in [37] of a cobordism which is a K^*-equivariant completion of the space

$$L(Y, D; X, 0) := O_Y(D) \cup_{U \times K^*} X \times (\mathbb{P}^1 \setminus \{0\}).$$

Another variant of our construction is given by Hu and Keel in [21].

2.4. Collapsibility

Definition 2.14 ([37]). Let X be a cobordism or any variety with a K^*-action.

1. We say that $F \in \mathcal{C}(X^{K^*})$ is an *immediate predecessor* of $F' \in \mathcal{C}(X^{K^*})$ if there exists a nonfixed point x such that $\lim_{t\to 0} tx \in F$ and $\lim_{t\to\infty} tx \in F'$.

2. We say that F *precedes* F' and write $F < F'$ if there exists a sequence of connected fixed point set components $F_0 = F, F_1, \ldots, F_l = F'$ such that F_{i-1} is an immediate predecessor of F_i (see [5]).

3. We call a cobordism (or a variety with a K^*-action) *collapsible* (see also Morelli [27]) if the relation $<$ on its set of connected components of the fixed point set is an order. (Here an order is just required to be transitive.)

Definition 2.15 ([3], [37]). A function $\chi\colon \mathcal{C}(X^{K^*}) \to \mathbb{Z}$ is *strictly increasing* if $\chi(F) < \chi(F')$ whenever $F < F'$.

2.5. Existence of a strictly increasing function for $\mathbb{P}^k$ and $\bar{B}$. The space $\mathbb{P}^k = \mathbb{P}(\mathbb{A}^{k+1})$ splits according to the weights as

$$\mathbb{P}^k = \mathbb{P}(\mathbb{A}^{k+1}) = \mathbb{P}(\mathbb{A}_{a_1} \oplus \cdots \oplus \mathbb{A}_{a_r})$$

where K^* acts on $\mathbb{A}_{a_i}$ with the weight a_i. Assume that $a_1 < \cdots < a_r$. Let $\bar{x}_{a_i} = [x_{i,1}, \ldots, x_{i,r_i}]$ be the coordinates on $\mathbb{A}_{a_i}$. The action of K^* is given by

$$t[\bar{x}_{a_1}, \ldots, \bar{x}_{a_r}] = [t^{a_1}\bar{x}_{a_1}, \ldots, t^{a_r}\bar{x}_{a_r}].$$

It follows that the fixed point components of $(\mathbb{P}^k)^{K^*}$ are $\mathbb{P}(\mathbb{A}_{a_i})$. We define a strictly increasing function $\chi_{\mathbb{P}}\colon \mathcal{C}(\mathbb{P}^{K^*}) \to \mathbb{Z}$ by

$$\chi_{\mathbb{P}}(\mathbb{P}(\mathbb{A}_{a_i})) = a_i.$$

We see that for $\bar{x} = [\bar{x}_{a_0}, \ldots, \bar{x}_{a_r}]$, $\lim_{t\to 0} tx \in \mathbb{P}(\mathbb{A}_{a_{\min}})$, $\lim_{t\to\infty} tx \in \mathbb{P}(\mathbb{A}_{a_{\max}})$, where

$$a_{\max} = \max\{a \mid \bar{x}_a \neq 0\}, \quad a_{\min} = \min\{a \mid \bar{x}_a \neq 0\}.$$

Then $\mathbb{P}(\mathbb{A}_{a_i}) < \mathbb{P}(\mathbb{A}_{a_j})$ iff $a_i < a_j$.

By the Sumihiro theorem ([33]), we embed $\bar{B}$ equivariantly into a projective space $\mathbb{P}^k$. Then every fixed point component F in $\mathcal{C}(\bar{B}^{K^*})$ is contained in $\mathbb{P}(\mathbb{A}_a) \in \mathcal{C}((\mathbb{P}^k)^{K^*})$ and we put $\chi_B(F) = \chi_{\mathbb{P}}(\mathbb{P}(\mathbb{A}_a)) = a_i$. The function $\chi_{\mathbb{P}}$ is strictly increasing on $\mathcal{C}((\mathbb{P}^k)^{K^*})$ and the function χ_B is strictly increasing on $\mathcal{C}(\bar{B}^{K^*})$. This implies

Lemma 2.16. *A compactified cobordism $\bar{B}$ is collapsible.*

2.6. Decomposition of a birational cobordism

Definition 2.17 ([3], [37]). A cobordism B is *elementary* if any $F \neq F' \in \mathcal{C}(B^{K^*})$ are incomparable with respect to $>$.

The function χ_F defines a decomposition of $\mathcal{C}(B^{K^*})$ into elementary cobordisms

$$B_{a_i} := B \setminus \Big(\bigcup_{\chi_B(F)<a_i} F^- \cup \bigcup_{\chi_B(F)>a_i} F^+ \Big),$$

where $a_1 < \cdots < a_r$ are the values of χ_B. This yields

Lemma 2.18. 1. $(B_{a_1})_- = B_-$, $(B_{a_r})_+ = B_+$.

2. $(B_{a_{i+1}})_- = (B_{a_i})_+ = B \setminus \big(\bigcup_{\chi_B(F)\le a_i} F^- \cup \bigcup_{\chi_B(F)\ge a_{i+1}} F^+\big)$.

3. $\chi(F) = a_i$ *for any* $F \in \mathcal{C}(B_{a_i})$.

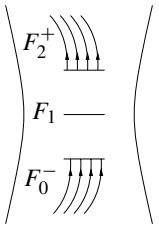


Figure 4. Elementary birational cobordism.

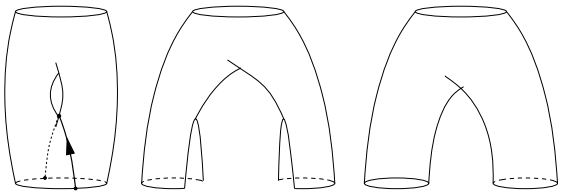


Figure 5. "Handle"-elementary cobordism in Morse Theory.

2.7. Decomposition of $\mathbb{P}^k$. Set $\mathbb{A}_{\geq a_i} := \mathbb{A}_{a_i} \oplus \cdots \oplus \mathbb{A}_{a_r}$, $\mathbb{A}_{>a_i} := \mathbb{A}_{a_{i+1}} \oplus \cdots \oplus \mathbb{A}_{a_r}$, and define $\mathbb{A}_{<a_i}$, $\mathbb{A}_{\leq a_i}$ analogously.

Lemma 2.19. $\mathbb{P}(\mathbb{A}_{a_i})^+ = \mathbb{P}(\mathbb{A}_{\geq a_i})$ *and* $\mathbb{P}(\mathbb{A}_{a_i})^- = \mathbb{P}(\mathbb{A}_{\leq a_i})$.

Lemma 2.20. *Set* $\mathbb{P}_{a_i} := \mathbb{P}^k \setminus \big(\bigcup_{\chi_{\mathbb{P}}(F)<a_i} F^- \cup \bigcup_{\chi_{\mathbb{P}}(F)>a_i} F^+\big)$. *Then*

$$\mathbb{P}_{a_i} = \mathbb{P}^k \setminus \mathbb{P}(\mathbb{A}_{>a_i}) \setminus \mathbb{P}(\mathbb{A}_{<a_i}), \quad (\mathbb{P}_{a_i})_+ = \mathbb{P}^k \setminus \mathbb{P}(\mathbb{A}_{\geq a_i}) \setminus \mathbb{P}(\mathbb{A}_{<a_i})$$
$$(\mathbb{P}_{a_i})_- = \mathbb{P}^k \setminus \mathbb{P}(\mathbb{A}_{>a_i}) \setminus \mathbb{P}(\mathbb{A}_{\leq a_i}).$$

Lemma 2.21. $B_{a_i} = \overline{B} \cap \mathbb{P}_{a_i}$, $(B_{a_i})_- = \overline{B} \cap (\mathbb{P}_{a_i})_-$, $(B_{a_i})_+ = \overline{B} \cap (\mathbb{P}_{a_i})_+$.

2.8. GIT and existence of quotients for $\mathbb{P}^k$. The sets $\mathbb{P}_{a_i}$ can be interpreted in terms of Mumford's GIT theory. Any lifting of the action of K^* on $\mathbb{P}^k = \mathbb{P}(\mathbb{A}^{k+1})$ to $\mathbb{A}^{k+1}$ is called a *linearization*. Consider the twisted action on $\mathbb{A}^{k+1}$,

$$t_r(x) = t^{-r} \cdot t(x).$$

The twisting does not change the action on $\mathbb{P}(\mathbb{A}^{k+1})$ and defines different linearizations. If we compose the action with a group monomorphism $t \mapsto t^k$ the weights of the new action $t^k(x)$ will be multiplied by k. The good and geometric quotients for $t(x)$ and $t^k(x)$ are the same. Keeping this in mind it is convenient to allow linearizations with rational weights.

Definition 2.22. A point $x \in \mathbb{P}^k$ is *semistable* with respect to t_r, written $x \in (\mathbb{P}^k, t_r)^{\mathrm{ss}}$, if there exists an invariant section $s \in \Gamma(\mathcal{O}_{\mathbb{P}^{k+1}}(n)^{t_r})$, for some $n \in \mathbb{N}$ such that $s(x) \neq 0$.

Lemma 2.23 ([3]). $\mathbb{P}_{a_i} = (\mathbb{P}^k, t_{a_i})^{\mathrm{ss}}$, $(\mathbb{P}_{a_i})_- = (\mathbb{P}^k, t_{a_i-\frac{1}{2}})^{\mathrm{ss}}$, $(\mathbb{P}_{a_i})_+ = (\mathbb{P}^k, t_{a_i+\frac{1}{2}})^{\mathrm{ss}}$.

Proof. $x \in \mathbb{P}_{a_i}$ iff either $\bar{x}_{a_i} \neq 0$ or $\bar{x}_{a_{j_1}} \neq 0$ and $\bar{x}_{a_{j_2}} \neq 0$ for $a_{j_1} < r = a_i < a_{j_2}$. In both situations we find a nonzero t_r-invariant section $s_i = x_i$ or $s_{j_1 j_2} = x_{j_1}^{b_1} x_{j_2}^{b_2}$ for suitable coprime b_1 and b_2.

$x \in (\mathbb{P}_{a_i})_-$ iff $\bar{x}_{a_{j_1}} \neq 0$ and $\bar{x}_{a_{j_2}} \neq 0$ for $a_{j_1} < a_i \leq a_{j_2}$ (or equivalently $a_{j_1} < r = a_i - 1/2 < a_{j_2}$). As before there is a nonzero t_r-invariant section $x_{j_1}^{b_1} x_{j_2}^{b_2}$ for suitable coprime b_1 and b_2. □

It follows from GIT theory that $(\mathbb{P}^k, t^r)^{\mathrm{ss}} /\!/ K^*$ exists and it is a projective variety. By Lemma 2.21 and the above we get

Corollary 2.24. *There exist quotients $\pi_{a_i} \colon B_{a_i} \to B_{a_i} /\!/ K^*$ and $\pi_{a_i-} = B_{a_i-} \to (B_{a_i})_-/K^*$, $\pi_{a_i+} = (B_{a_i})_+ \to (B_{a_i})_+/K^*$.*

2.9. Local description

Proposition 2.25 ([37]). *Let B_a be a smooth elementary cobordism. Then for any $x \in F_0$ there exists an invariant neighborhood V_x of x and a K^*-equivariant étale morphism (i.e. locally analytic isomorphism) $\phi \colon V_x \to \mathrm{Tan}_x$, where $\mathrm{Tan}_x \simeq \mathbb{A}_K^n$ is the tangent space with the induced linear K^*-action, such that in the diagram*

$$\begin{array}{ccccccc}
(B_a)_-/K^* & \supset & V_x /\!/ K^* \times_{\mathrm{Tan}_x /\!/ K^*} \mathrm{Tan}_{x-}/K^* & \simeq & V_{x-}/K^* & \to & \mathrm{Tan}_{x-}/K^* \\
\downarrow & & & & \downarrow & & \downarrow \\
B_a /\!/ K^* & & \supset & & V_x /\!/ K^* & \to & \mathrm{Tan}_x /\!/ K^* \\
\uparrow & & & & \uparrow & & \uparrow \\
(B_a)_+/K^* & \supset & V_x /\!/ K^* \times_{\mathrm{Tan}_x /\!/ K^*} \mathrm{Tan}_{x+}/K^* & \simeq & V_{x+}/K^* & \to & \mathrm{Tan}_{x+}/K^*
\end{array}$$

the vertical arrows are defined by open embeddings and the horizontal morphisms are defined by ϕ and are étale.

Proof. By taking local semiinvariant parameters at the point $x \in F_0$ one can construct an equivariant morphism $\phi \colon U_x \to \mathrm{Tan}_x \simeq \mathbb{A}_K^n$ from some open affine invariant neighborhood U_x such that ϕ is étale at x. By Luna's Lemma (see [Lu], Lemme 3 (Lemme Fondamental)) there exists an invariant affine neighborhood $V_x \subseteq U_x$ of the point x such that $\phi_{|V_x}$ is étale, the induced map $\phi_{|V_x /\!/ K^*} \colon V_x /\!/ K^* \to \mathrm{Tan}_x /\!/ K^*$ is étale and $V_x \simeq V_x /\!/ K^* \times_{\mathrm{Tan}_x /\!/ K^*} \mathrm{Tan}_x$. This defines the isomorphisms $V_x /\!/ K^* \times_{\mathrm{Tan}_x /\!/ K^*} \mathrm{Tan}_{x-}/K^* \simeq V_{x-}/K^*$. Note that $(B_a)_- = B_a \setminus \bigcup F^+_{F \in \mathcal{C}(B^{K^*})}$ and $V_x \cap F^+ = (V_x \cap F)^+$. (Both sets are closed and irreducible.) Thus $(V_x)_- = V_x \cap (B_a)_-$ and we get the horizontal inclusions. □

Proposition 2.26 ([37]). *There is a factorization of the morphism $\phi \colon Y \to X$ given by $Y = (B_{a_1})_-/K^* \dashrightarrow (B_{a_1})_+/K^* = (B_{a_2})_-/K^* \dashrightarrow \cdots \dashrightarrow (B_{a_{k-1}})_+/K^* = (B_{a_k})_-/K^* \dashrightarrow (B_{a_k})_+/K^* = X$.*

Remark 2.27. The birational maps $(B_a)_-/K^* \dashrightarrow (B_a)_+/K^*$ are locally described by Example 2.7. Both spaces have cyclic singularities and differ by the composite of a weighted blow-up and a weighted blow-down. To achieve the factorization we need to desingularize quotients as in for instance case 1 of the example. It is hopeless to modify weights by birational modification of smooth varieties. Instead we want to view Example 2.7 from the perspective of toric varieties.

3. Toric varieties

3.1. Fans and toric varieties. Let $N \simeq \mathbb{Z}^k$ be a lattice contained in the vector space $N^{\mathbb{Q}} := N \otimes \mathbb{Q} \supset N$.

Definition 3.1 ([11], [31]). By a *fan* Σ in $N^{\mathbb{Q}}$ we mean a finite collection of finitely generated strictly convex cones σ in $N^{\mathbb{Q}}$ such that

- any face of a cone in Σ belongs to Σ,
- any two cones of Σ intersect in a common face.

If σ is a face of σ' we shall write $\sigma \preceq \sigma'$.

We say that a cone σ in $N^{\mathbb{Q}}$ is *regular* if it is generated by a part of a basis of the lattice $e_1, \ldots, e_k \in N$, written $\sigma = \langle e_1, \ldots, e_k \rangle$. A cone σ is *simplicial* if it is generated over $\mathbb{Q}$ by linearly independent integral vectors $v_1, \ldots, v_k$, written $\sigma = \langle v_1, \ldots, v_k \rangle$

Definition 3.2. Let Σ be a fan and $\tau \in \Sigma$. The *star* of the cone τ and the *closed star* of τ are defined as follows:

$$\mathrm{Star}(\tau, \Sigma) := \{\sigma \in \Sigma \mid \tau \preceq \sigma\},$$

$$\overline{\mathrm{Star}}(\tau, \Sigma) := \{\sigma \in \Sigma \mid \sigma' \preceq \sigma \text{ for some } \sigma' \in \mathrm{Star}(\tau, \Sigma)\}.$$

To a fan Σ there is associated a toric variety $X_\Sigma \supset T$, i.e. a normal variety on which a torus T acts effectively with an open dense orbit (see [23], [12], [31], [15]). To each cone $\sigma \in \Sigma$ corresponds an open affine invariant subset X_σ and its unique closed orbit O_σ. The orbits in the closure of the orbit O_σ correspond to the cones of $\mathrm{Star}(\sigma, \Sigma)$. In particular, $\tau \preceq \sigma$ iff $\overline{O}_\tau \supset O_\sigma$.

The fan Σ is *nonsingular* (resp. *simplicial*) if all its cones are nonsingular (resp. simplicial). Nonsingular fans correspond to nonsingular varieties.

Denote by

$$M := \mathrm{Hom}_{\mathrm{alg.gr.}}(T, K^*)$$

the lattice of group homomorphisms to K^*, i.e. characters of T. The dual lattice $\mathrm{Hom}_{\mathrm{alg.gr.}}(K^*, T)$ of 1-parameter subgroups of T can be identified with the lattice N. Then the vector space $M^{\mathbb{Q}} := M \otimes \mathbb{Q}$ is dual to $N^{\mathbb{Q}} = N \otimes \mathbb{Q}$.

The elements $F \in M = N^*$ are functionals on N and integral functionals on $N^{\mathbb{Q}}$. For any $\sigma \subset N^{\mathbb{Q}}$ we denote by

$$\sigma^\vee := \{F \in M \mid F(v) \geq 0 \text{ for any } v \in \sigma\}$$

the set of integral vectors of the dual cone to σ. Then the ring of regular functions $K[X_\sigma]$ is $K[\sigma^\vee]$.

We call a vector $v \in N$ *primitive* if it generates the sublattice $\mathbb{Q}_{\geq 0} v \cap N$. Primitive vectors correspond to 1-parameter monomorphisms.

For any $\sigma \subset N^{\mathbb{Q}}$ set

$$\sigma^\perp := \{m \in M \mid (v, m) = 0 \text{ for any } v \in \sigma\}.$$

The latter set represents all invertible characters on X_σ. All noninvertible characters are in $\sigma^\vee \setminus \sigma^\perp$ and vanish on O_σ. The ring of regular functions on $O_\sigma \subset X_\sigma$ can be written as $K[O_\sigma] = K[\sigma^\perp] \subset K[\sigma^\vee]$.

3.2. Star subdivisions and blow-ups

Definition 3.3 ([23], [31], [12], [15]). A *birational toric morphism* or simply a *toric morphism* of toric varieties $X_\Sigma \to X_{\Sigma'}$ is a morphism identical on $T \subset X_\Sigma, X_{\Sigma'}$.

By the *support* of a fan Σ we mean the union of all its faces, $|\Sigma| = \bigcup_{\sigma \in \Sigma} \sigma$.

Definition 3.4 ([23], [31], [12], [15]). A *subdivision* of a fan Σ is a fan Δ such that $|\Delta| = |\Sigma|$ and any cone $\sigma \in \Sigma$ is a union of cones $\delta \in \Delta$.

Definition 3.5. Let Σ be a fan and ϱ be a ray passing in the relative interior of $\tau \in \Sigma$. Then the *star subdivision* $\varrho \cdot \Sigma$ of Σ with respect to ϱ is defined to be

$$\varrho \cdot \Sigma = (\Sigma \setminus \mathrm{Star}(\tau, \Sigma)) \cup \{\varrho + \sigma \mid \sigma \in \overline{\mathrm{Star}}(\tau, \Sigma) \setminus \mathrm{Star}(\tau, \Sigma)\}.$$

If Σ is nonsingular, i.e. all its cones are nonsingular, $\tau = \langle v_1, \ldots, v_l \rangle$ and $\varrho = \langle v_1 + \cdots + v_l \rangle$ then we call the star subdivision $\varrho \cdot \Sigma$ *nonsingular*.

Proposition 3.6 ([23], [12], [31], [15]). *Let X_Σ be a toric variety. There is a* 1-1 *correspondence between subdivisions of the fan Σ and proper toric morphisms $X_{\Sigma'} \to X_\Sigma$.*

Remark 3.7. Nonsingular star subdivisions from 3.5 correspond to blow-ups of smooth varieties at closures of orbits ([31], [15]). Arbitrary star subdivisions correspond to blow-ups of some ideals associated to valuations (see Lemma 5.20).

4. Polyhedral cobordisms of Morelli

4.1. Preliminaries. By $N^{\mathbb{Q}+}$ we shall denote a vector space $N^{\mathbb{Q}+} \approx \mathbb{Q}^k$ containing a lattice $N^+ \simeq \mathbb{Z}^k$, together with a primitive vector $v_0 \in N^+$ and the canonical projection

$$\pi : N^{\mathbb{Q}+} \to N^{\mathbb{Q}} \simeq N^{\mathbb{Q}+}/\mathbb{Q} \cdot v_0.$$

Definition 4.1 ([27]). A cone $\sigma \subset N^{\mathbb{Q}+}$ is *π-strictly convex* if $\pi(\sigma)$ is strictly convex (contains no line). A fan Σ is π-strictly convex if it consists of π-strictly convex cones.

In the following all the cones in $N^{\mathbb{Q}+}$ are assumed to be π-strictly convex and simplicial. The π-strictly convex cones σ in $N^{\mathbb{Q}+}$ split into two categories.

Definition 4.2. A cone $\sigma \subset N^{\mathbb{Q}+}$ is called *independent* if the restriction of π to σ is a linear isomorphism (equivalently $v_0 \notin \operatorname{span}(\sigma)$). A cone $\sigma \subset N^{\mathbb{Q}+}$ is called *dependent* if the restriction of π to σ is a lattice submersion which is not an isomorphism (equivalently $v_0 \in \operatorname{span}(\sigma)$).

A dependent cone is called a *circuit* if all its proper faces are independent.

Lemma 4.3. *Any dependent cone σ contains a unique circuit δ.*

4.2. K^*-actions and $N^{\mathbb{Q}+}$. The vector $v_0 = (a_1, \dots, a_k) \in N^{\mathbb{Q}+}$ defines a 1-parameter subgroup $t^{v_0} := t_1^{a_1} \dots t_k^{a_k}$ acting on T and all toric varieties $X \supset T$. Denote by M^+ the lattice dual to N^+. Then the lattice $N := N^+/\mathbb{Z} \cdot v_0$ is dual to the lattice $M := \{a \in M^+ \mid (a, v_0) = 0\}$ of all the characters invariant with respect to the group action. The natural projection of cones $\pi : \sigma \to \sigma^\Gamma$ defines the good quotient morphism

$$X_\sigma = \operatorname{Spec} K[\sigma^\vee] \to X_\sigma // K^* = \operatorname{Spec} K[\sigma^\vee \cap M] = \operatorname{Spec} K[(\sigma^\Gamma)^\vee] = X_{\sigma^\Gamma}.$$

Lemma 4.4. *A cone σ is independent iff the geometric quotient $X_\sigma \to X_\sigma/K^*$ exists or alternatively if X_σ contains no fixed points. The cone σ is dependent if O_σ is a fixed point set.*

Proof. Note that the set $X_\sigma^{K^*}$ is closed and if it is nonempty then it contains O_σ. Then a point $p \in O_\sigma$ is fixed, i.e. $t^{v_0} p = p$, iff for all functionals $F \in \sigma^\perp$ (i.e. $x^F(p) \neq 0$) we have $x^F(p) = x^F(t^{v_0} p) = t^{F(v_0)} x^F(p)$.

Then for all $F \in \sigma^\perp \subset \operatorname{span}(\sigma)^\perp$ we have $F(v_0) = 0$ so $v_0 \in \operatorname{span}(\sigma)$. □

Corollary 4.5. *A cone $\delta \in \Sigma$ is a circuit if and only if O_δ is the generic orbit of some $F \in \mathcal{C}(X_\Sigma^{K^*})$.*

Proof. O_σ is fixed with respect to the action of K^* if σ is dependent. Thus $O_\sigma \subset \overline{O}_\delta$ where δ is the unique circuit in σ (Lemma 4.3). □

4.3. Morelli cobordisms

Definition 4.6 (Morelli [27], [4]). A fan Σ in $N^{\mathbb{Q}+} \supset N^+$ is called a *polyhedral cobordism* or simply a cobordism if the sets of cones

$$\partial_-(\Sigma) := \{\sigma \in \Sigma \mid \text{there is } p \in \operatorname{int}(\sigma) \text{ so that } p - \varepsilon \cdot v_0 \notin |\Sigma| \text{ for all small } \varepsilon > 0\},$$

$$\partial_+(\Sigma) := \{\sigma \in \Sigma \mid \text{there is } p \in \operatorname{int}(\sigma) \text{ so that } p + \varepsilon \cdot v_0 \notin |\Sigma| \text{ for all small } \varepsilon > 0\}$$

are subfans of Σ and $\pi(\partial_-(\Sigma)) := \{\pi(\tau) \mid \tau \in \partial_-(\Sigma)\}$ and $\pi(\partial_+(\Sigma)) := \{\pi(\tau) \mid \tau \in \partial_+(\Sigma)\}$ are fans in $N^{\mathbb{Q}}$.

4.4. Dependence relation. Let $\sigma = \langle v_1, \dots, v_k \rangle$ be a dependent (simplicial) cone. Then, by definition $v_0 \in \operatorname{span}(v_1, \dots, v_k)$ where $v_1, \dots, v_k$ are linearly independent. There exists a unique up to rescaling integral relation

$$r_1 v_1 + \cdots + r_k v_k = a v_0, \quad \text{where } a > 0. \tag{$*$}$$

Definition 4.7 ([27]). The rays of σ are called *positive, negative* and *null* vectors, according to the sign of the coefficient in the defining relation.

Remark 4.8. Note that the relation $(*)$ defines a unique relation

$$r'_1 w_1 + \cdots + r'_k w_k = 0 \tag{$**$}$$

where w_i are generating vectors in the rays $\pi(\langle v_i \rangle)$, $r'_i w_i = r_i \pi(v_i)$. In particular $r'_i / r_i > 0$.

Lemma 4.9. *Let $\sigma = \langle v_1, \dots, v_k \rangle$ be a dependent cone. Then an independent face τ is in $\partial_+(\sigma)$ (resp. $\tau \in \partial_+(\sigma)$) if τ is a face of $\langle v_1, \dots, \check{v}_i, \dots, v_k \rangle$ for some index i such that $r_i < 0$ (resp. $r_i > 0$).*

Proof. By definition $\tau \in \partial_+(\sigma)$ there exists $p \in \operatorname{int}(\tau)$ such that for any sufficiently small $\varepsilon > 0$, $p + \varepsilon v_0 \notin \sigma$. Write $p = \sum \alpha_i v_i = \sum_{r_i > 0} \alpha_i v_i + \sum_{r_i < 0} \alpha_i v_i + \sum_{r_i = 0} \alpha_i v_i$, where $\alpha_i \geq 0$. Then one of the coefficients in

$$p + \varepsilon v_0 = \sum_{r_i > 0} (\alpha_i + r_i \varepsilon) v_i + \sum_{r_i < 0} (\alpha_i + r_i \varepsilon) v_i + \sum_{r_i = 0} (\alpha_i + r_i \varepsilon) v_i$$

is negative for small $\varepsilon > 0$. This is possible if $\alpha_i = 0$ for some index i with $r_i < 0$. □

Lemma 4.10. *A cone τ is in $\partial_+(\sigma)$ iff there exists $F \in \sigma^\vee \cap \tau^\perp$ such that $F(v_0) < 0$.*

Proof. If $\tau \in \partial_+(\sigma)$ then there exists $p \in \operatorname{int}(\tau)$ for which $p + \varepsilon v_0 \notin \sigma$. Hence there exists $F \in \sigma^\vee$ such that $F(p + \varepsilon v_0) < 0$ for small $\varepsilon > 0$. Then $F(p) = 0$ and $F(v_0) < 0$. Since $p \in \operatorname{int}(\tau)$ we have $F_{|\tau} = 0$. □

Corollary 4.11. *$\partial_+(\sigma)$ (resp. $\partial_-(\sigma)$) is a fan.*

Proof. By the lemma above, if $\tau \in \sigma^+$ then every face τ' of τ is in σ^+. □

Lemma 4.12. *Let σ be a dependent cone in $N^{\mathbb{Q}+}$. Then $B := X_\sigma$ is a birational cobordism such that*

- $(X_\sigma)_+ = X_{\partial_-(\sigma)}$, $(X_\sigma)_- = X_{\partial_+(\sigma)}$.
- $(X_\sigma)_+ / K^* \cong X_{\pi(\partial_-(\sigma))}$, $(X_\sigma)_- / K^* \cong X_{\pi(\partial_+(\sigma))}$.
- *$\pi(\partial_-(\sigma))$ and $\pi(\partial_+(\sigma))$ are both decompositions of $\pi(\sigma)$.*
- *There is a factorization into a sequence of proper morphisms $(X_\sigma)_+ / K^* \to (X_\sigma) /\!/ K^* \leftarrow (X_\sigma)_- / K^*$.*

Proof. We have $p \in O_\tau$ where $O_\tau \subset (X_\sigma)_-$ iff $\lim t^{v_0} p \notin X_\sigma$. This is equivalent to existence of a functional $F \in \sigma^\vee$ for which $x^F(t^{v_0} p) = t^{F(v_0)} x^F(p)$ has a pole at $t = 0$. This means exactly that $x^F(p) \neq 0$ and $F(v_0) < 0$. The last condition says $F_{|\tau} = 0$ and $F(v_0) < 0$, which is equivalent to $\tau \in \partial_+(\sigma)$.

Suppose that $x \in \pi(\sigma)$. Then $\pi^{-1}(x) \cap \sigma$ is a line segment or a point. Let $y = \sup\{\pi^{-1}(x) \cap \sigma\}$. Then $y \in \operatorname{int}(\tau)$, where $\tau \prec \sigma$ and $y + \varepsilon v_0 \notin \sigma$, which implies that $\tau \in \partial_+(\sigma)$. Thus every point in $\pi(\sigma)$ belongs to a relative interior of a unique cone $\pi(\tau) \in \pi(\partial_+(\sigma))$. Since $\pi_{|\tau}$ is a linear isomorphism and $\partial_+(\sigma)$ is a fan, all faces of $\pi(\tau)$ are in $\pi(\partial_-(\sigma))$. Finally, $\pi(\partial_+(\sigma))$ and $\pi(\partial_-(\sigma))$ are both decompositions of $\pi(\sigma)$ corresponding to toric varieties $(X_\sigma)_-/K^* = X_{\pi(\partial_+(\sigma))}$ and $(X_\sigma)_+/K^* = X_{\pi(\partial_-(\sigma))}$. □

The above yields

Lemma 4.13. *$B = X_\sigma$ is an elementary cobordism with a single fixed point component $F := \overline{O}_\delta$, where $\delta = \langle v_i \mid r_i \neq 0\rangle$ is a circuit. Moreover $(X_\sigma)_+ = X_{\partial_-(\sigma)} = X_\sigma \setminus \overline{O}_{\sigma_+}$, where*

$$\sigma_+ := \langle v_i \mid r_i > 0\rangle, \quad \sigma_- := \langle v_i \mid r_i < 0\rangle.$$

In particular $F^+ = (\overline{O}_\delta)^+ = \overline{O}_{\sigma_+}$ and $F^- = (\overline{O}_\delta)^- = \overline{O}_{\sigma_-}$.

4.5. Example 2.7 revisited. The cobordism X_σ from the lemma generalizes the cobordism $B = \mathbb{A}^{l+m+r}_K \supset T = (K^*)^{l+m+r}$ from Example 2.7. The action of K^* determines a 1-parameter subgroup of T which corresponds to a vector $v_0 = [a_1, \ldots, a_l, -b_1, \ldots, -b_m, 0, \ldots, 0]$. The cobordism B is associated with a nonsingular cone $\Delta \subset N_{\mathbb{Q}}$, while B_- and B_+ correspond to the fans $\partial_+(\Delta)$ and $\partial_-(\Delta)$ consisting of the faces of Δ visible from above and below respectively.

The quotients B_+/K^*, B_-/K^* and $B/\!/K^*$ are toric varieties corresponding to the fans $\pi(\partial_+(\Delta)) = \{\pi(\sigma) \mid \sigma \in \partial_+(\Delta)\}$, $\pi(\partial_-(\Delta)) = \{\pi(\sigma) \mid \sigma \in \Delta^-\}$ and $\pi(\Delta)$ respectively, where π is the projection defined by v_0.

The relevant birational map $\phi\colon B_-/K^* \dashrightarrow B_+/K^*$ for $l, m \geq 2$ is a toric flip associated with a bistellar operation replacing the triangulation $\pi(\partial_-(\Delta))$ of the cone $\pi(\Delta)$ with $\pi(\partial_+(\Delta))$.

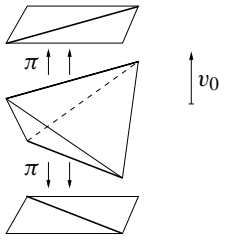


Figure 6. Morelli cobordism

4.6. π-nonsingular cones

Definition 4.14 (Morelli). An independent cone τ is *π-nonsingular* if $\pi(\tau)$ is nonsingular. A fan Σ is *π-nonsingular* if all independent cones in Σ are π-nonsingular. In particular a dependent cone σ is *π-nonsingular* if all its independent faces are π-nonsingular.

Lemma 4.15. *Let $\sigma = \langle v_1, \dots, v_k\rangle$ be a dependent cone and w_i be primitive generators of the rays $\pi(v_i)$. Let $\sum r_i' w_i = 0$ be the unique relation $(**)$ between vectors w_i. Then the ray $\varrho := \pi(\sigma_+) \cap \pi(\sigma_-)$ is generated by the vector $\sum_{r_i'>0} r_i' w_i = \sum_{r_i'<0} -r_i' w_i$ and $\varrho \cdot \pi(\partial_+(\sigma)) = \varrho \cdot \pi(\partial_-(\sigma))$. If σ is a π-nonsingular dependent cone then the ray ϱ defines regular star subdivisions of $\pi(\partial_+(\sigma))$ and $\pi(\partial_-(\sigma))$.*

Proof. Note that $\pi(\partial_+(\sigma)) \setminus \pi(\partial_-(\sigma))$ are exactly the cones containing $\pi(\sigma_+)$. That is, $\pi(\partial_+(\sigma)) \setminus \pi(\partial_-(\sigma)) = \mathrm{Star}(\pi(\sigma_+), \pi(\partial_+(\sigma)))$. This gives $\varrho \cdot \pi(\partial_+(\sigma)) = (\pi(\sigma_+) \cap \pi(\sigma_-)) \cup \{\varrho + \tau \mid \tau \in \pi(\sigma_+) \cap \pi(\sigma_-)\} = \varrho \cdot \pi(\sigma_-)$. Assume now that σ is π-nonsingular and all the coefficients r_i' are coprime. By Lemma 4.9 and the π-nonsingularity the set of vectors $w_1, \dots, \check{w}_i, \dots, w_k$ where $r_i' \neq 0$ is a basis of the lattice $\pi(\sigma) \cap N$. Thus every vector w_i, where $r_i' \neq 0$, can be written as an integral combination of others. Since the relation $(**)$ is unique it follows that the coefficient r_i' is equal to ± 1. Thus ϱ is generated by the vector $\sum_{r_i'>0} w_i = \sum_{r_i'<0} w_i$ and determines regular star subdivisions. □

Corollary 4.16. *If σ is dependent then there exists a factorization*

$$(X_\sigma)_-/K^* \xleftarrow{\phi_-} \Gamma((X_\sigma)_-/K^*, (X_\sigma)_+/K^*) \xrightarrow{\phi_+} (X_\sigma)_+/K^*,$$

where $\Gamma((X_\sigma)_-/K^*, (X_\sigma)_+/K^*)$ is the normalization of the graph of $(X_\sigma)_-/K^* \to (X_\sigma)_+/K^*$. If σ is π-nonsingular the morphisms ϕ_-, ϕ_+ are blow-ups of smooth centers.

Proof. By definition $\Gamma((X_\sigma)_-/K^*, (X_\sigma)_+/K^*)$ is a toric variety. By the universal property of the graph (dominating component of the fiber product) it corresponds to the coarsest simultaneous subdivision of both $\pi(\sigma_-)$ and $\pi(\sigma_+)$, that is, to the fan $\{\tau_1 \cap \tau_2 \mid \tau_1 \in \pi(\sigma_-), \tau_2 \in \pi(\sigma_+)\} = \varrho \cdot \pi(\sigma_-) = \varrho \cdot \pi(\sigma_+)$. □

4.7. The π-desingularization lemma of Morelli and centers of blow-ups.

For any simplicial cone $\sigma = \langle v_1, \dots, v_k\rangle$ in N set

$$\mathrm{par}(\sigma) := \{v \in \sigma \cap N_\sigma \mid v = \alpha_1 v_1 + \cdots + \alpha_k v_k, \text{ where } 0 \le \alpha_i < 1\},$$

$$\overline{\mathrm{par}(\sigma)} := \{v \in \sigma \cap N_\sigma \mid v = \alpha_1 v_1 + \cdots + \alpha_k v_k, \text{ where } 0 \le \alpha_i \le 1\}.$$

We associate with a dependent cone σ and an integral vector $v \in \pi(\sigma)$ a vector $\mathrm{Mid}(v, \sigma) := \pi^{-1}_{|\partial_-(\sigma)}(v) + \pi^{-1}_{|\partial_+(\sigma)}(v) \in \sigma$ ([27]), where $\pi_{|\partial_-(\sigma)}$ and $\pi_{|\partial_+(\sigma)}$ are the restrictions of π to $\partial_-(\sigma)$ and $\partial_+(\sigma)$.

We also set $\mathrm{Ctr}_-(\sigma) := \sum_{r_i<0} w_i$, $\mathrm{Ctr}_+(\sigma) := \sum_{r_i>0} w_i$.

Lemma 4.17 (Morelli [27], [28], [4]). *Let Σ be a simplicial cobordism in N^+. Then there exists a simplicial cobordism Δ obtained from Σ by a sequence of star subdivisions such that Δ is π-nonsingular. Moreover, the sequence can be taken so that any independent and already π-nonsingular face of Σ remains unaffected during the process. All the centers of the star subdivisions are of the form $\pi_{|\tau}^{-1}(\mathrm{par}(\pi(\tau)))$ where τ is independent, and* $\mathrm{Mid}(\mathrm{Ctr}_{\pm}(\sigma), \sigma)$*, where σ is dependent.*

Remark 4.18. It follows from Lemma 4.17 that π-desingularization can be done for an open affine neighborhood of a point x of $F \in \mathcal{C}(B^{K^*})$ on the smooth cobordism B which is étale isomorphic with the tangent space Tan_x. We need to show how to globalize this procedure in a coherent and possibly canonical way. This will replace the tangent space Tan_x in the local description of flips defined by elementary cobordisms (as in Proposition 2.25) with π-nonsingular X_σ.

By Corollary 4.16 we get a factorization into a blow-up and a blow-down at smooth centers: $(B_a)_-/K^* \xleftarrow{\phi_-} \Gamma((B_a)_-/K^*, (B_a)_+/K^*) \xrightarrow{\phi_+} (B_a)_+/K^*$.

5. π-desingularization of birational cobordisms

5.1. Stratification by isotropy groups on a smooth cobordism. Let B be a smooth cobordism of dimension n. Denote by Γ_x the isotropy group of a point $x \in B$. Define the stratum $s = s_x$ through x to be an irreducible component of the set $\{p \in B \mid \Gamma_x = \Gamma_p\}$.

We can find Γ_x-semiinvariant parameters in the affine open neighborhood U of x such that Γ_x acts nontrivially on $u_1, \ldots, u_k$ and trivially on $u_{k+1}, \ldots, u_n$.

After suitable shrinking of U the parameters define an étale Γ_x-equivariant morphism $\varphi\colon U \to \mathrm{Tan}_x = \mathbb{A}^n$. By definition the stratum s is locally described by $u_1 = \cdots = u_k = 0$. The parameters $u_1, \ldots, u_k$ determine a Γ_x-equivariant smooth morphism

$$\psi\colon U \to \mathrm{Tan}_{B,x}/\mathrm{Tan}_{s,x} = \mathbb{A}^k.$$

We shall view $\mathbb{A}^k = X_\sigma$ as a toric variety with a torus T_σ and refer to ψ as a *toric chart*. This assigns to a stratum s the cone σ and the relevant group Γ_σ acting on X_σ. Then Luna's [24] fundamental lemma implies that the morphisms ϕ and ψ preserve stabilizers, the induced morphism $\psi_\Gamma\colon U/\!/\Gamma_x \to X_\sigma/\!/\Gamma_\sigma$ is smooth and $U \simeq U/\!/\Gamma_x \times_{\mathbb{A}^k/\!/\Gamma_x} \mathbb{A}^k$.

The invariant Γ_x can be defined for $X_\sigma = \mathbb{A}^k$ and determine the relevant T_σ-invariant stratification S_σ on X_σ. By shrinking U we may assume that the strata on U are inverse images of the strata on X_σ. Any stratum s_y on U through y after a suitable rearrangement of $u_1, \ldots, u_k$ is described in the neighborhood $U' \subset U$ of y by $u_1 = \cdots = u_\ell = 0$, where $\Gamma_y \leq \Gamma_x$ acts nontrivially on $u_1, \ldots, u_\ell$ and trivially on $u_{\ell+1} \ldots, u_k, u_{k+1}, \ldots, u_n$. The remaining Γ_y-invariant parameters at y are $u_{\ell+1} - u_{\ell+1}(y), \ldots, u_n - u_n(y)$. Then the closure of $\bar{s}_y$ is described on U by $u_1 = \cdots = u_\ell = 0$ and contains s_x. This shows

Lemma 5.1. *The closure of any stratum is a union of strata.*

We can introduce an order on the strata by setting

$$s' \leq s \quad \text{iff} \quad \bar{s}' \subseteq s.$$

Lemma 5.2. *If $s' \leq s$ then there exists an inclusion $i_{\sigma'\sigma} : \sigma' \hookrightarrow \sigma$ onto a face of σ. The inclusion $i_{\sigma'\sigma}$ defines a $\Gamma_{\sigma'}$-equivariant morphism of toric varieties $X_{\sigma'} \to X_{\sigma'} \times 1 \hookrightarrow X_{\sigma'} \times T \subset X_\sigma$, where $T_{\sigma'} \times T = T_\sigma$ and $\Gamma_{\sigma'} \subset T_{\sigma'}$. Moreover we can write $X_\sigma \cong X_{\sigma'} \times \mathbb{A}^r$ where $\Gamma_{\sigma'}$ acts trivially on $\mathbb{A}^r$ and nontrivially on all coordinates of $X_{\sigma'} \simeq \mathbb{A}^\ell$.*

In the above situation we shall write

$$\sigma' \leq \sigma.$$

The lemma above immediately implies

Lemma 5.3. *If $\tau < \sigma$ (that is, $\tau \leq \sigma$, $\tau \neq \sigma$) then $\Gamma_\tau \subsetneq \Gamma_\sigma$.*

Consider the stratification S_σ on X_σ. Every stratum $s_\tau \in S_\sigma$, where $\tau \leq \sigma$, is a union of orbits $O_{\tau'}$. Set

$$\bar{\tau} := \{\tau' \mid O_{\tau'} \subset s_\tau\}.$$

Lemma 5.4. *Any cone from the set $\tau' \in \bar{\tau}$ can be expressed as $\tau' \simeq \tau \times \langle e_1, \ldots, e_r \rangle \subset \sigma$, and $X_{\tau'} = X_\tau \times \mathbb{A}^s \times T^{r-s}$ where Γ_τ acts trivially on $\mathbb{A}^r \times T^{r-s}$.*

Lemma 5.5. *We have $\Gamma_\tau = \Gamma_{\tau'} := \{g \in \Gamma_\sigma \mid \text{for all } x \in O_{\sigma'},\ gx = x\}$ for any $\tau' \in \bar{\tau}$.*

5.2. Local projections

Definition 5.6. A cone σ in $N^{\mathbb{Q}}$ is *of maximal dimension* if $\dim \sigma = \dim N^{\mathbb{Q}}$.

Every cone σ in $N^{\mathbb{Q}}$ defines a cone of maximal dimension in $N^{\mathbb{Q}} \cap \operatorname{span}\{\sigma\}$ with lattice $N \cap \operatorname{span}\{\sigma\}$. We denote it by $\underline{\sigma}$. There is a noncanonical isomorphism

$$X_\sigma = X_{\underline{\sigma}} \times O_\sigma.$$

The vector space $\operatorname{span}\{\sigma\} \subset N^{\mathbb{Q}}$ corresponds to a subtorus $T_{\underline{\sigma}} \subset T_\sigma$ defined as $T_{\underline{\sigma}} := \{t \in T_\sigma \mid tx = x \text{ for } x \in O_\sigma\}$. Then O_σ is isomorphic to the torus $T_\sigma / T_{\underline{\sigma}}$ with dual lattice $\sigma^\perp \subset M^{\mathbb{Q}}$.

Lemma 5.7. *If $\Gamma \subset T_\sigma$ acts freely on $X_\sigma = X_{\underline{\sigma}} \times O_\sigma$ then*

$$X_\sigma / \Gamma = X_{\underline{\sigma}} \times O_\sigma / \Gamma,$$

where $O_\sigma \simeq O_\sigma / \Gamma$ if Γ is finite, while O_σ / Γ is isomorphic to a torus of dimension $\dim O_\sigma - 1$ if $\Gamma = K^*$.

Proof. By assumption $\Gamma \cap T_{\underline{\sigma}}$ is trivial. Hence Γ acts trivially on $X_{\underline{\sigma}}$ and $X_\sigma/\Gamma = X_{\underline{\sigma}} \times O_\sigma/\Gamma$. □

Let $\pi_\sigma : (\sigma, N_\sigma) \to (\sigma^\Gamma, N_\sigma^\Gamma)$ denote the projection corresponding to the quotient map $X_\sigma \to X_\sigma//\Gamma_\sigma$.

Lemma 5.8. *If $\tau \le \sigma$ then $\pi_\tau(\tau) \simeq \pi_\sigma(\tau)$.*

Proof. $X_{\underline{\tau}} \times O_\tau$ is an open subvariety in X_σ and Γ_τ acts trivially on O_τ. We have

$$(X_{\underline{\tau}} \times O_\tau)/\Gamma_\tau = X_{\underline{\tau}}/\Gamma_\tau \times O_\tau = X_{\pi_\tau(\tau)} \times O_\tau.$$

$\Gamma_\sigma/\Gamma_\tau$ acts freely on $(X_{\underline{\tau}} \times O_\tau)/\Gamma_\tau = X_{\pi_\tau(\tau)} \times O_\tau$. Thus by the previous lemma $X_{\pi_\sigma(\tau)} \cong X_{\pi_\tau(\tau)} \times O_\tau/\Gamma_\sigma$. □

Lemma 5.9. *Let Γ be a subgroup of Γ_σ, and $\pi_\Gamma : \sigma \to \sigma^\Gamma$ be the projection corresponding to the quotient $X_\sigma \to X_\sigma/\Gamma$. For any $\tau \le \sigma$ and $\tau' \in \bar{\tau}$ we have $\tau' = \tau \oplus \langle e_1, \dots, e_k\rangle$ where $\langle e_1, \dots, e_k\rangle$ is regular and $\pi_\Gamma(\tau') = \pi_\Gamma(\tau) \oplus \langle e_1, \dots, e_k\rangle$.*

Proof. $X_{\tau'} = X_\tau \times \mathbb{A}^k \times O_{\tau'}$ where the action of $\Gamma_\tau \cap \Gamma$ on $\mathbb{A}^k \times O_\tau$ is trivial. Thus $X_{\tau'}/\Gamma_\tau = X_\tau/\Gamma_\tau \times \mathbb{A}^k \times O_{\tau'}$. Now $\Gamma/(\Gamma_\tau \cap \Gamma)$ acts freely on $O_{\tau'} \subset s_\tau$ and we use Lemma 5.7. □

5.3. Independent and dependent cones. By Lemma 5.8 there exists a lattice isomorphism $j_{\tau\sigma} : \pi_\tau(\tau) \to \pi_\sigma(\tau)$, where $\tau \le \sigma$. Thus the projections π_τ and π_σ are coherent and related: $j_{\tau\sigma}\pi_\tau = \pi_\sigma$.

Case 1: $\Gamma_\sigma = K^*$. The action of K^* on X_σ corresponds to a primitive vector $v_\sigma \in N_\sigma$. The invariant characters $M_\sigma^\Gamma \subset M_\sigma$ are precisely those $F \in M_\sigma^\Gamma$ such that $F(v_\sigma) = 0$. The dual morphism is a projection $\pi_\sigma : N_\sigma \to N_\sigma/\mathbb{Z} \cdot v_\sigma = N_\sigma^\Gamma$.

The quotient morphism of toric varieties $X_\sigma \to X_\sigma/\Gamma_\sigma$ corresponds to the projection $\sigma \to \pi_\sigma(\sigma)$.

Case 2: $\Gamma_\sigma \cong \mathbb{Z}_n$. The invariant characters $M_\sigma^\Gamma \subset M_\sigma$ form a sublattice of dimension $\dim(M_\sigma^\Gamma) = \dim(M_\sigma)$, where $M_\sigma/M_\sigma^\Gamma \simeq \mathbb{Z}_n$. The dual morphism defines an inclusion $\pi : N_\sigma \hookrightarrow N_\sigma^\Gamma$. The projection $\sigma \to \pi_\sigma(\sigma)$ is a linear isomorphism which does not preserve lattices. This gives

Lemma 5.10. *X_τ is independent iff Γ_τ is finite. X_σ is dependent iff $\Gamma_\sigma = K^*$.*

Definition 5.11. Let Δ^σ be a decomposition of a cone $\sigma \in \Sigma$. A cone $\tau \in \Delta^\sigma$ is *independent* if $\pi_{\sigma|\tau}$ is a linear isomorphism. A cone τ is *dependent* if $\pi_{\sigma|\tau}$ is not a linear isomorphism.

5.4. Semicomplexes and birational modification of cobordisms. By glueing cones σ corresponding to strata along their faces we construct a *semicomplex* Σ, that is, a partially ordered set of cones such that for $\sigma \leq \sigma'$ there exists a face inclusion $i_{\sigma\sigma'}\colon \sigma \to \sigma'$.

Remark 5.12. The glueing need not be transitive: for $\sigma \leq \sigma' \leq \sigma''$ we have $i_{\sigma'\sigma''}i_{\sigma\sigma'} \neq i_{\sigma\sigma''}$. Instead, there exists an automorphism α_σ of σ such that $i_{\sigma'\sigma''}i_{\sigma\sigma'} = i_{\sigma\sigma''}\alpha_\sigma$.

For any fan Σ denote by $\mathrm{Vert}(\Sigma)$ the set of all 1-dimensional faces (rays) in Σ. Denote by $\mathrm{Aut}(\sigma)$ the automorphisms of σ inducing Γ_σ-equivariant automorphisms.

Definition 5.13. By a *subdivision* of Σ we mean a collection $\Delta = \{\Delta^\sigma \mid \sigma \in \Sigma\}$ of subdivisions Δ^σ of σ such that:

1. If $\tau \leq \sigma$ then the restriction $\Delta^\sigma_{|\tau}$ of Δ^σ to τ is equal to Δ^τ.
2. All rays in $\mathrm{Vert}(\Delta^\sigma) \setminus \mathrm{Vert}(\sigma)$ are contained in $\bigcup_{\tau\leq\sigma} \mathrm{int}(\tau)$.
3. Δ^σ is $\mathrm{Aut}(\sigma)$-invariant.

Remark 5.14. Condition 3 is replaced with a stronger one in the following proposition.

Lemma 5.15. *If $\tau' \in \bar{\tau}$, $\tau' \prec \sigma \in \Sigma$ then* $\mathrm{Vert}(\Delta^\sigma_{|\tau'}) \setminus \mathrm{Vert}(\tau') \subset \tau$ *and*

$$\Delta^\sigma_{|\tau'} = \Delta^\sigma_{|\tau} \oplus \langle e_1, \ldots, e_k\rangle = \Delta^\tau \times \langle e_1, \ldots, e_k\rangle.$$

Lemma 5.16. *For every point $x \in B \setminus (B_+ \cap B_-)$, $x \in s'$ there exists a toric chart $x \in U_\sigma \to X_\sigma$, with $\Gamma_\sigma = K^*$, corresponding to a stratum $s \subset \bar{s}'$. In particular the maximal cones of Σ are circuits.*

Proof. Let τ correspond to a stratum $s' \ni x$. By definition of cobordism $\lim_{t\to 0} tx = x_0$ or $\lim_{t\to\infty} tx = x_0$ exists. The point x_0 is K^*-fixed and belongs to a stratum s, with $\Gamma_s = \Gamma_\sigma = K^*$. Since U is a K^*-invariant neighborhood of x_0 it contains an orbit $K^* \cdot x$ and the point x. Moreover $\bar{s}' \supset s$ and $\tau \leq \sigma$. □

Lemma 5.17. *Let σ be the cone corresponding to a stratum s on B and $x \in s$. Then* $\widehat{X}_x = \mathrm{Spec}\,\widehat{O}_{x,B} \simeq (X_\sigma \times \mathbb{A}^{\dim(s)})^\wedge \cong \mathrm{Spec}\,K[[x_1, \ldots, x_k, \ldots, x_n]]$.

Set $\widetilde{X}_\sigma := (X_\sigma \times \mathbb{A}^{\dim(s)})^\wedge$ and let G_σ denote the group of all Γ_σ-equivariant autorphisms of $\widetilde{X}_\sigma$.

The subdivision Δ^σ of σ defines a toric morphism and induces a proper birational Γ_σ-equivariant morphism

$$\widetilde{X}_{\Delta^\sigma} := X_{\Delta^\sigma} \times_{X_\sigma} \widetilde{X}_\sigma \to \widetilde{X}_\sigma.$$

Proposition 5.18. *Let $\Delta = \{\Delta^\sigma \mid \sigma \in \Sigma\}$ be a subdivision of Σ such that:*

$$\text{For every } \sigma \in \Sigma \text{ the morphism } \widetilde{X}_{\Delta^\sigma} \to \widetilde{X}_\sigma \text{ is } G_\sigma\text{-equivariant.} \tag{1}$$

Then Δ defines a K^*-equivariant birational modification $f\colon B' \to B$ such that for every toric chart $\varphi_\sigma\colon U \to X_\sigma$ there exists a Γ_σ-equivariant fiber square

$$\begin{array}{rcl} U_\sigma \times_{X_\sigma} X_{\Delta^\sigma} \simeq f^{-1}(U_\sigma) & \to & X_{\Delta^\sigma} \\ \downarrow f & & \downarrow \\ U_\sigma & \to & X_\sigma. \end{array} \tag{2}$$

Definition 5.19. A decomposition Δ of Σ is *canonical* if it satisfies condition (1).

Proof. The above diagrams define open subsets $f_\sigma^{-1}(U_\sigma)$ together with proper birational Γ_σ-equivariant morphisms $f_\sigma^{-1}(U_\sigma) \to U_\sigma$. Let $s' \le s$ be a stratum corresponding to the cone $\tau \le \sigma$. By Lemma 5.15, the restriction of the diagram (2) defined by $U_\sigma \to X_\sigma$ to a neighborhood U_τ of $y \in s'$ determines a diagram defined by the induced toric chart $U_\tau \to X_\tau$ and the decomposition Δ^τ of τ. In order to show that the $f_\sigma^{-1}(U)$ glue together we need to prove that for $x \in s$ and two different charts of the form $\varphi_{1,\sigma}\colon U_{1,\sigma} \to X_\sigma$ and $\varphi_{2,\sigma}\colon U_{2,\sigma} \to X_\sigma$ where $x \in U_{1,\sigma}, U_{2,\sigma}$ the induced varieties $V_1 := f_{1,\sigma}^{-1}(U_{1,\sigma})$ and $V_2 := f_{2,\sigma}^{-1}(U_{2,\sigma})$ are isomorphic over $U_{1,\sigma} \cap U_{2,\sigma}$. For simplicity assume that $U_{1,\sigma} = U_{2,\sigma} = U$ by shrinking $U_{1,\sigma}$ and $U_{2,\sigma}$ if necessary. The charts $\varphi_{1,\sigma}, \varphi_{2,\sigma}\colon U \to X_\sigma$ are defined by the two sets of semiinvariant parameters, $u_1^1, \dots, u_k^1$ and $u_1^2, \dots, u_k^2$ with a nontrivial action of Γ_σ. These sets can be extended to full sets of parameters $u_1^1, \dots, u_k^1, u_{k+1}, \dots, u_n$ and $u_1^2, \dots, u_k^2, u_{k+1}, \dots, u_n$ where Γ_σ acts trivially on $u_{k+1}, \dots, u_n$, and $u_{k+1} \dots, u_n$ define parameters on the stratum s at x. These two sets of parameters define étale morphisms $\varphi_{1,\sigma}, \varphi_{2,\sigma}\colon U \to X_\sigma \times \mathbb{A}^{n-k}$ and fiber squares

$$\begin{array}{rcl} \bar{\varphi}_{i,\sigma}\colon\ V_i & \to & X_{\Delta^\sigma} \times \mathbb{A}^{n-k} \\ \downarrow & & \downarrow \\ \varphi_{i,\sigma}\colon\ U & \to & X_\sigma \times \mathbb{A}^{n-k}. \end{array}$$

Suppose the induced Γ-equivariant birational map $f\colon V_1 \dashrightarrow V_2$ is not an isomorphism over U.

Let V be the graph of f which is a dominating component of the fiber product $V_1 \times_U V_2$. Then either $V \to V_1$ or $V \to V_2$ is not an isomorphism (i.e. collapses a curve to a point) over some $x \in s \cap U$. Consider an étale Γ_σ-equivariant morphism $e\colon \widehat{X}_x \to U$. Pull-backs of the morphisms $V_i \to U$ via e define two different nonisomorphic Γ_σ-equivariant liftings $Y_i \to \widehat{X}_x$, since the graph Y of $Y_1 \dashrightarrow Y_2$ (which is a pull-back of V) is not isomorphic to at least one Y_i. But these two liftings are defined by two isomorphisms $\widehat{\varphi}_1, \widehat{\varphi}_2\colon \widehat{X}_x \simeq \widetilde{X}_\sigma$. These isomorphisms differ by some automorphism $g \in G_\sigma$, so we have $\widehat{\varphi}_1 = g \circ \widehat{\varphi}_2$. Since g lifts to the automorphism of $\widetilde{X}_{\Delta^\sigma}$ we get $Y_1 \simeq Y_2 \simeq \widetilde{X}_{\Delta^\sigma}$, which contradicts the choice of Y_i.

Thus V_1 and V_2 are isomorphic over any $x \in s$ and B' is well defined by glueing pieces $f_\sigma^{-1}(U)$ together. We need to show that the action of K^* on B lifts to the action of K^* on B'.

Note that B' is isomorphic to B over the open generic stratum $U \supset B_+ \cup B_-$ of points x with $\Gamma_x = \{e\}$. By Lemma 5.16 every point $x \in B \setminus (B_+ \cap B_-)$ is in U_σ, with $\Gamma_\sigma = K^*$. Then the diagram (2) defines the action of K^* on $f^{-1}(U_\sigma)$. □

5.5. Basic properties of valuations. Let $K(X)$ be the field of rational functions on an algebraic variety or an integral scheme X. A *valuation* on $K(X)$ is a group homomorphism $\mu\colon K(X)^* \to G$ from the multiplicative group $K(X)^*$ to a totally ordered group G such that $\mu(a+b) \geq \min(\mu(a), \mu(b))$. By the *center* of a valuation μ on X we mean an irreducible closed subvariety $Z(\mu) \subset X$ such that for any open affine $V \subset X$, intersecting $Z(\mu)$, the ideal $I_{Z(\mu)\cap V} \subset K[V]$ is generated by all $f \in K[V]$ such that $\mu(f) > 0$ and for any $f \in K[V]$, we have $\mu(f) \geq 0$. Each vector $v \in N^{\mathbb{Q}}$ defines a linear function on M which determines a valuation $\mathrm{val}(v)$ on a toric variety $X_\Sigma \supset T$.

For any regular function $f = \sum_{w\in M} a_w x^w \in K[T]$ set

$$\mathrm{val}(v)(f) := \min\{(v, w) \mid a_w \neq 0\}.$$

If $v \in \mathrm{int}(\sigma)$, where $\sigma \in \Delta$, then $\mathrm{val}(v)$ is positive for all x^F, where $F \in \sigma^\vee \setminus \sigma^\perp$. In particular we get

$$Z(\mathrm{val}(v)) = \overline{O}_\sigma \quad \text{iff} \quad v \in \mathrm{int}\,\sigma.$$

If $v \in \sigma$ then $\mathrm{val}(v)$ is a valuation on $R = K[X_\sigma] = K[\sigma^\vee]$, that is, $\mathrm{val}(v)(f) \geq 0$ for all $f \in K[\sigma^\vee] \setminus \{0\}$. We construct ideals for all $a \in \mathbb{N}$ which uniquely determine $\mathrm{val}(v)$:

$$I_{\mathrm{val}(v),a} = \{f \in R \mid \mathrm{val}(v)(f) \geq a\} = (x^F \mid F \in \sigma^\vee, F(v) \geq a) \subset R.$$

By glueing $I_{\mathrm{val}(v),a}$ for all $v \in \sigma$ and putting $\mathcal{I}_{\mathrm{val}(v),a|X_\sigma} = \mathcal{O}_{X_\sigma}$ if $v \notin \sigma$ we construct a coherent sheaf of ideals $\mathcal{I}_{\mathrm{val}(v),a}$ on X_Δ.

Lemma 5.20 ([23]). *The star subdivision $\langle v\rangle \cdot \Sigma$ corresponds to the normalized blow-up of $\mathcal{I}_{\mathrm{val}(v),a}$ on X_Σ for a sufficiently divisible $a \in \mathbb{N}$.*

5.6. Stable vectors. Let $g\colon X \to Y$ be any dominant morphism of integral schemes (that is, $\overline{g(X)} = Y$) and μ be a valuation of $K(X)$. Then g induces a valuation $g_*(\mu)$ on $K(Y) \simeq g(K(Y)) \subset K(X)$: $g_*\mu(f) = \mu(f \circ g)$.

Definition 5.21. Let Σ be the semicomplex defined for the cobordism B. A vector $v \in \mathrm{int}(\sigma)$, where $\sigma \in \Sigma$, is called *stable* if for every $\sigma \leq \sigma'$, $\mathrm{val}(v)$ is $G_{\sigma'}$-invariant on $\widetilde{X}_{\sigma'}$.

Lemma 5.22. *If $\widetilde{X}_{\Delta^\sigma} \to \widetilde{X}_\sigma$ is G_σ-equivariant and* $\mathrm{val}(v)$ *is G_σ-invariant then $\widetilde{X}_{\langle v\rangle\cdot\Delta^\sigma} \to \widetilde{X}_\sigma$ is G_σ-equivariant.*

Proof. The morphism $\widetilde{X}_{\langle v\rangle\cdot\Delta^\sigma} \to \widetilde{X}_{\Delta^\sigma}$ is a pull-back of the morphism $X_{\langle v\rangle\cdot\Delta^\sigma} \to X_{\Delta^\sigma}$. Thus, by Lemma 5.20, $\widetilde{X}_{\langle v\rangle\cdot\Delta^\sigma} \to \widetilde{X}_{\Delta^\sigma}$ is a normalized blow-up of $\mathcal{I}_{\mathrm{val}(v),a}$ on $\widetilde{X}_{\Delta^\sigma}$. But the latter sheaf is G_σ-invariant. □

Proposition 5.23. *Let $\Delta = \{\Delta^\sigma \mid \sigma \in \Sigma\}$ be a canonical subdivision of Σ and v be a stable on Σ. Then $\langle v\rangle \cdot \Delta := \{\langle v\rangle \cdot \Delta^\sigma \mid \sigma \in \Sigma\}$ is a canonical subdivision of Σ.*

5.7. Convexity

Lemma 5.24. *Let* $\mathrm{val}(v_1)$ *and* $\mathrm{val}(v_2)$ *be G_σ-invariant valuations on X_σ. Then all valuations* $\mathrm{val}(v)$*, where $v = av_1 + bv_2$, $a, b \geq 0$, $a, b \in \mathbb{Q}$, are G_σ-invariant.*

Proof. Let $\Delta = \langle v_1\rangle \cdot \langle v_2\rangle \cdot \sigma$ be a subdivision of σ. Then by Lemma 5.22, the morphism $\widetilde{X}_\Delta \to \widetilde{X}_\sigma$ is G_σ-equivariant. The exceptional divisors D_1 and D_2 of the morphism are G_σ-invariant and correspond to one-dimensional cones (rays) $\langle v_1\rangle, \langle v_2\rangle \in \Delta$. The cone $\tau = \langle v_1, v_2\rangle \in D$ corresponds to the orbit O_τ whose closure is $D_1 \cap D_2$ and thus the generic point is G_σ-invariant. The action of G_σ on $\widetilde{X}_\sigma$ induces an action on the local ring $\widetilde{X}_{\Delta,O_\tau}$ at the generic point of O_τ and on its completion $K(O_\tau)[[\underline{\tau}^\vee]]$. Note that for any $v \in \tau$, $\mathrm{val}(v)_{|K(O_\tau)} = 0$. For any $F \in \underline{\tau}^\vee = \frac{\tau^\vee}{\tau^\perp}$ the divisor (x^F) of the character x^F on $\widehat{X}_\tau := \operatorname{Spec} K(O_\tau)[[\underline{\tau}^\vee]]$ is a combination $n_1D_1 + n_2D_2$ for $n_1n_2 \in \mathbb{Z}$. Since D_1 and D_2 are G_σ-invariant, the divisor $(x^F) = n_1D_1 + n_2D_2$ is G_σ-invariant, that is, for any $g \in G$, we have $gx^F = u_{g,F} \cdot x^F$ where $u_{g,F}$ is invertible on $K(O_\tau)[[\underline{\tau}^\vee]]$. Thus for every $v \in \tau$ and $g \in G$ we have

$$\begin{aligned} g^*(I_{\mathrm{val}(v),a}) &= g^*(x^F | F \in \underline{\tau}^\vee, F(v) \geq a) \\ &= (u_{g,F}x^F | F \in \underline{\tau}^\vee, F(v) \geq a) = I_{\mathrm{val}(v),a}. \end{aligned}$$

Thus $\mathrm{val}(v)$ is G_σ-invariant on $K(O_\tau)[[\underline{\tau}^\vee]]$ and on its subring $\mathcal{O}_{\widetilde{X}_{\Delta,O_\tau}}$. The latter ring has the same quotient field as $\widetilde{X}_\sigma$ so $\mathrm{val}(v)$ is G_σ-invariant on $\widetilde{X}_\sigma$. □

Lemma 5.25. *Let $\sigma \in \Sigma$ and $v_1, v_2 \in \sigma$ be stable vectors. Then all vectors $v = av_1 + bv_2 \in \sigma$, where $a, b \in \mathbb{Q}_{>0}$, are stable.*

5.8. Basic properties of stable vectors

Lemma 5.26. *Let* $\mathrm{Tan}_0 = \mathbb{A}^n = \mathrm{Tan}_0^{a_0} \oplus \mathrm{Tan}_0^{a_1} \oplus \cdots \oplus \mathrm{Tan}_0^{a_k}$ *denote the tangent space of* $\widetilde{X}_\sigma = \operatorname{Spec} K[[u_1, \ldots, u_n]]$ *at* 0 *and its decomposition according to the weight distribution. Let $d\colon G_\sigma \to \mathrm{GL}(\mathrm{Tan}_0)$ be the differential morphism defined as $g \mapsto dg$. Then $d(G_\sigma) = \mathrm{GL}(\mathrm{Tan}_0^{a_1}) \times \cdots \times \mathrm{GL}(\mathrm{Tan}_0^{a_k})$.*

Lemma 5.27. *Let $v \in \sigma$, where $\sigma \in \Sigma$, be an integral vector such that for any $g \in G_\sigma$, there exists an integral vector $v_g \in \sigma$ such that $g_*(\mathrm{val}(v)) = \mathrm{val}(v_g)$. Then* $\mathrm{val}(v)$ *is G_σ-invariant on $\widetilde{X}_\sigma$.*

Proof. Set $W = \{v_g \mid g \in G\}$. For any natural number n, the ideals $I_{\mathrm{val}(v_g),a}$ are generated by monomials. They define the same Hilbert–Samuel function $k \mapsto \dim_K(K[\widetilde{X}_\sigma]/(I_{\mathrm{val}(v_g),a} + m^k))$, where $m \subset K[\widetilde{X}_\sigma]$ denotes the maximal ideal. It follows that the set W is finite. On the other hand since $I_{\mathrm{val}(v_g),a}$ are generated by monomials they are uniquely determined by the ideals $\mathrm{gr}(I_{\mathrm{val}(v_g),a})$ in the graded ring

$$\mathrm{gr}(O_{\widetilde{X}_\sigma}) = O_{\widetilde{X}_\sigma}/m \oplus m/m^2 \oplus \cdots .$$

The connected group $d(G_\sigma)$ acts algebraically on $\mathrm{gr}(O_{\widetilde{X}_\sigma})$ and on the connected component of the Hilbert scheme with fixed Hilbert polynomial. In particular it acts trivially on its finite subset W and consequently $d(G_\sigma)$ preserves $\mathrm{gr}(I_{\mathrm{val}(v_g),a})$ and G_σ preserves $I_{\mathrm{val}(v_g),a}$. □

Let $R \subset K$ be a ring contained in the field. We can order valuations by writing

$$\mu_1 > \mu_2 \quad \text{if} \quad \mu_1(a) \geq \mu_2(a) \text{ for all } a \in R \text{ and } \mu_1 \neq \mu_2.$$

A cone σ defines a partial ordering: $v_1 > v_2$ if $v_1 - v_2 \in \sigma$. Both orders coincide for $K[X_\sigma] \subset K(X_\sigma)$: $v_1 > v_2$ iff $\mathrm{val}(v_1) > \mathrm{val}(v_2)$.

Lemma 5.28. *Let σ be a cone in $N_\sigma^{\mathbb{Q}}$ with the lattice of 1-parameter subgroups $N_\sigma \subset N_\sigma^Q$ and the dual lattice of characters M_σ. Let μ be any integral (or rational) valuation centered on $\overline{O}_\tau$, where $\tau \preceq \sigma$. Then the restriction of μ to $M_\sigma \subset K(\widetilde{X}_\sigma)^*$ defines a functional on $\tau^\vee \subseteq M_\sigma^Q$ corresponding to a vector $v_\mu \in \operatorname{int}\tau$ such that $F(v_\mu) = \mu(x^F)$ for $F \in M_\sigma$ and $\mu \geq \mathrm{val}(v_\mu)$ on $\widetilde{X}_\sigma$.*

Proof. $I_{\mu,a} \supseteq (x^F \mid \mu(x^F) \geq a) = (x^F \mid F(v_\mu) \geq a) = I_{\mathrm{val}(v_\mu),a}$. □

Lemma 5.29. *Let $\Gamma \subset \Gamma_\sigma$ be a finite group acting on $\widetilde{X}_\sigma$. Let $\pi : N^{\mathbb{Q}} \to (N^\Gamma)^{\mathbb{Q}}$ denote the projection corresponding to the geometric quotient $\widetilde{X}_\sigma \to \widetilde{X}_{\pi(\sigma)} = \widetilde{X}_\sigma/\Gamma$. Then* $\mathrm{val}(v)$ *is G_σ-invariant on $\widetilde{X}_\sigma$ iff* $\mathrm{val}(\pi(v))$ *is G_σ-invariant on $\widetilde{X}_{\pi(\sigma)}$.*

Proof. ($\Rightarrow$) $\mathrm{val}(v)$ is G_σ-invariant on $K[\widetilde{X}_\sigma]$ and it is invariant on $K[\widetilde{X}_\sigma]^\Gamma$.

($\Leftarrow$) Note that π defines an inclusion of same dimension lattices $N \hookrightarrow N^\Gamma$ and $M^\Gamma \hookrightarrow M$.

Assume that $\mathrm{val}(\pi(v))$ is G_σ-invariant. It defines a functional on the lattice M^Γ and its unique extension to $M \supset M^\Gamma$ corresponding to $\mathrm{val}(v)$. Since $g_*(\mathrm{val}(\pi(v))) = \mathrm{val}(\pi(v))$, we have $g_*(\mathrm{val}(v))_{|M^\Gamma} = \mathrm{val}(v)_{|M^\Gamma}$ and consequently $g_*(\mathrm{val}(v))_{|M} = \mathrm{val}(v)_{|M}$. By Lemma 5.28, $g_*(\mathrm{val}(v)) \geq \mathrm{val}(v)$ for all $g \in G_\sigma$. Thus $\mathrm{val}(v) \geq g_*^{-1}(\mathrm{val}(v))$ for all $g^{-1} \in G_\sigma$. Finally $g_*(\mathrm{val}(v)) = \mathrm{val}(v)$. □

5.9. Stability of centers from $\mathrm{par}(\pi(\tau))$. In the following let Δ^σ be a decomposition of $\sigma \in \Sigma$ such that $\widetilde{X}_{\Delta^\sigma} \to \widetilde{X}_\sigma$ is G_σ-equivariant, $\tau \in \Delta^\sigma$ be its face and Γ be a finite subgroup of Γ_σ. Denote by $\pi : (\sigma, N_\sigma) \to (\sigma^\Gamma, N_\sigma^\Gamma)$ the linear isomorphism and the lattice inclusion corresponding to the quotient $X_\sigma \to X_\sigma/\Gamma = X_{\pi(\sigma)}$.

Lemma 5.30. *Assume that for any $g \in G_\sigma$, there exists a cone $\tau_g \in \Delta^\sigma$ such that $g \cdot (\overline{O}_\tau) = \overline{O}_{\tau_g}$. Let $v \in \mathrm{int}(\pi(\tau)) \cap N_\sigma^\Gamma$ be an integral vector such that* val(v) *is not G_σ-invariant on $\widetilde{X}_\sigma/\Gamma$. Then there exist integral vectors $v_1 \in \mathrm{int}(\pi(\tau))$ and $v_2 \in \pi(\tau)$ such that*

$$v = v_1 + v_2.$$

Moreover if there exists $v_0 \in \pi(\sigma)$ (not necessarily integral) such that val(v_0) *is G_σ-invariant and $v > v_0$ on $\pi(\sigma)$ then $v_1 > v_0$ on $\pi(\sigma)$.*

Proof. If val(v) is not G_σ-invariant on $\widetilde{X}_\sigma/\Gamma$ then by Lemma 5.27 there exists an element $g \in G_\sigma$ such that $\mu_g := g_*(\mathrm{val}(v))$ is not a toric valuation. By the assumption μ_g is centered on $\overline{O}_{\pi(\tau_g)}$. Then by Lemma 5.28 it defines $v_g \in \mathrm{int}\,\pi(\tau_g)$ such that $\mu_g(x^F) = F(v_g)$ for $F \in \sigma^\vee$. Moreover $\mu_g > \mathrm{val}(v_g)$. Then the valuation $g_*^{-1}(\mathrm{val}(v_g))$ is centered on $\overline{O}_{\pi(\tau)}$. Thus it defines an integral $v_1 \in \mathrm{int}(\pi(\tau))$ such that $v \geq v_1$ on $\pi(\tau)$ and $v_2 := v - v_1$. Then

$$\mathrm{val}(v) = g_*^{-1}(\mu_g) > g_*^{-1}(\mathrm{val}(v_g)) \geq \mathrm{val}(v_1).$$

Note also that if $v \geq v_0$ then $\mu_g = g_*(\mathrm{val}(v)) \geq \mathrm{val}(v_0)$ and $\mathrm{val}(v_g) \geq \mathrm{val}(v_0)$. Thus also $\mathrm{val}(v_1) \geq \mathrm{val}(v_0)$. □

Lemma 5.31. *All valuations* val(v)*, where $v \in \varrho$, $\varrho \in \mathrm{Vert}(\Delta^\sigma) \setminus \mathrm{Vert}(\sigma)$, are G_σ-invariant.*

Proof. Let v_ϱ be the primitive generator of $\varrho \in \mathrm{Vert}(\Delta^\sigma) \setminus \mathrm{Vert}(\sigma)$. The ray ϱ corresponds to an exceptional divisor D_ϱ. By the definition there is no decomposition $v_\varrho = v_1 + v$. Thus by the previous lemma (for $\Gamma = \{e\}$), val(v) is G_σ-invariant. □

Lemma 5.32. *For any $\tau \leq \sigma$, the closure of the orbit $\overline{O}_\tau \subset \widetilde{X}_\sigma$ is G_σ-invariant.*

Proof. By Lemma 5.2, the ideal of $\overline{O}_\tau \subset \widetilde{X}_\sigma$ is generated by all functions with nontrivial Γ_σ-weights. □

Lemma 5.33. *The valuations* val(v)*, where $v \in \mathrm{par}(\pi(\tau))$, are G_σ-invariant on $\widetilde{X}_{\Delta^\sigma}$. Moreover $v \in \mathrm{int}(\pi(\sigma_0))$, for some $\sigma_0 \leq \sigma$.*

Proof. Let $v \in \mathrm{par}(\pi(\tau))$, where $\pi(\tau) \in \pi(\Delta)$ is a minimal integral vector such that val(v) is not G_σ-invariant. We may assume that $v \in \mathrm{int}(\pi(\tau))$ passing to its face if necessary. Let $\sigma' \in \bar{\sigma}_0$ be a face of σ such that $v \in \mathrm{int}\,\pi(\sigma')$. In particular $\pi(\sigma') \subset \pi(\tau)$. Then $\pi(\Delta^\sigma)_{|\pi(\sigma')} = \pi(\Delta^\sigma)_{|\pi(\sigma_0)} \oplus \langle e_1, \ldots, e_k \rangle$ by Lemmas 5.9 and 5.15 and $v \in \mathrm{par}(\pi(\tau)) \subset \pi(\sigma_0)$. Thus $\sigma' = \sigma_0$ and $v \in \mathrm{int}(\pi(\sigma_0))$. Let

$$\pi(\tau) = \langle v_1, \ldots, v_k, w_1, \ldots, w_\ell \rangle,$$

where $v_1, \ldots, v_k \in \mathrm{Vert}(\pi(\tau))$ and $w_1, \ldots, w_\ell \in \mathrm{Vert}(\pi(\Delta)) \setminus \mathrm{Vert}(\pi(\sigma))$. By Lemma 5.31, $\mathrm{val}(w_1), \ldots, \mathrm{val}(w_\ell)$ are G_σ-invariant. Write

$$v = \alpha_1 v_1 + \cdots + \alpha_k v_k + \alpha_{k+1} w_1 + \cdots + \alpha_{k+\ell} w_\ell,$$

where $0 < \alpha_i < 1$. Note that

$$v \geq v_0 = \alpha_{k+1} w_1 + \cdots + \alpha_{k+\ell} w_\ell$$

and $\overline{O}_{\pi(\sigma_0)} \subset \widetilde{X}_{\pi(\sigma)}$ is G_σ-invariant. By Lemma 5.30 for $v \in \pi(\sigma_0) \leq \pi(\sigma)$ and $v > v_0$ we can find integral vectors $v', v'' \in \pi(\sigma)$ such that $v = v' + v''$, $v' \geq v_0$. Then

$$v'' := v - v' \leq v - v_0 = \alpha_1 v_1 + \cdots + \alpha_k v_k.$$

Thus $v'' \in \mathrm{par}\langle v_1, \ldots, v_k\rangle \subseteq \mathrm{par}(\pi)(\tau)$. Write $v'' := \beta_1 v_1 + \cdots + \beta_k v_k$, where $\beta_i \leq \alpha_i$. Then

$$v' = v - v'' = (\alpha_1 - \beta_1)v_1 + \cdots + (\alpha_k - \beta_k)v_k + \alpha_{k+1} w_1 + \cdots + \alpha_{k+\ell} \in \mathrm{par}(\pi(\tau)).$$

By the minimality assumption, $\mathrm{val}(v')$ and $\mathrm{val}(v'')$ are G_σ-invariant and it follows from Lemma 5.24 that $\mathrm{val}(v) = \mathrm{val}(v' + v'')$ is G_σ-invariant. □

Corollary 5.34. *Let $\Delta = \{\Delta^\sigma \in \Sigma\}$ be a decomposition of Σ. Let $\tau \in \Delta^\sigma$ be an independent face. Then the vectors in $(\pi_{\sigma_{|\tau}})^{-1}(\mathrm{par}(\pi_\sigma(\tau)))$ are stable.*

Proof. Put $\Gamma = \Gamma_\tau$. Let $\pi : (\sigma, N_\sigma) \to (\sigma, N_\sigma^\Gamma)$ be the linear isomorphism and a lattice inclusion corresponding to the quotient $X_\sigma \to X_\sigma / \Gamma$. Then by Lemma 5.8, $\pi(\tau) \simeq \pi_\tau(\tau) \simeq \pi_\sigma(\tau)$ and by Lemma 5.33 vectors in $(\pi_{\sigma_{|\tau}})^{-1}(\mathrm{par}(\pi_\sigma(\tau))) = \pi^{-1}(\mathrm{par}(\pi(\tau)))$ are stable. □

Corollary 5.35. 1. *Assume that for any $g \in G_\sigma$, there exists $\tau_g \in \Delta^\sigma$ such that $g(\overline{O}_\tau) = \overline{O}_{\tau_g}$. Then $\overline{O}_\tau$ is G_σ-invariant. Moreover all valuations* $\mathrm{val}(v)$*, where $v \in \overline{\mathrm{par}}\,(\tau) \cap \mathrm{int}(\tau)$, are G_σ-invariant.*

2. *Let $\tau \in \Delta^\sigma$ be an independent cone such that $\overline{O}_\tau$ is G_σ-invariant. Then for any $v \in \pi_\sigma^{-1}(\overline{\mathrm{par}}\,(\pi(\tau)) \cap \mathrm{int}(\pi(\tau)))$ the valuation* $\mathrm{val}(v)$ *is G_σ-invariant.*

Proof. 1. Let $\tau = \langle v_1, \ldots, v_k\rangle$ and $v = \alpha_1 v_1 + \cdots + \alpha_k v_k$, where $0 < \alpha_i \leq 1$, be a minimal vector in $\mathrm{int}(\tau) \cap \overline{\mathrm{par}}\,(\tau)$ such that $\mathrm{val}(v)$ is not G_σ-invariant. Then by Lemma 5.30, the vector v can be written as $v = v' + v''$, where $v', v'' < v$, $v' \in \mathrm{int}(\tau)$, $v'' \in \tau$. Thus $v' = \alpha_1' v_1 + \cdots + \alpha_k' v_k$ where $0 < \alpha_i' \leq \alpha_i \leq 1$ and $v'' = \alpha_1'' v_1 + \cdots + \alpha_k'' v_k$, where $0 \leq \alpha_i'' = \alpha_i - \alpha_i' < 1$. Then $v' \in \mathrm{int}(\tau) \cap \overline{\mathrm{par}}\,(\tau)$ and $v'' \in \mathrm{par}(\tau)$. By Corollary 5.34, $\mathrm{val}(v'')$ is G_σ-invariant on $\widetilde{X}_\sigma$. By the minimality assumption $\mathrm{val}(v')$ is G_σ-invariant. Since $v = v' + v''$, the valuation $\mathrm{val}(v)$ is G_σ-invariant on $\widetilde{X}_\sigma$ and its center $Z(\mathrm{val}(v))$ equals $\overline{O}_\tau$.

2. Let $\pi : N \to N^\Gamma$ be the projection corresponding to the quotient $X_\sigma \to X_\sigma / \Gamma_\tau$. Then $\pi(\tau) \simeq \pi_\sigma(\tau)$. The proof is now exactly the same as the proof in 1 except that we replace $\widetilde{X}_{\Delta^\sigma}$ with $\widetilde{X}_{\Delta^\sigma}/\Gamma_\tau$. □

Corollary 5.36. *Let $\delta \in \Delta^\sigma$ be a circuit. Then $\overline{O}_\delta$ is G_σ-invariant.*

Proof. By Corollary 4.5, $\overline{O}_\delta$ is an irreducible component of a G_σ-invariant closed subscheme $\widetilde{X}_{\Delta^\sigma}^{K^*}$. Thus by the previous corollary it is G_σ-invariant. □

5.10. Stability of $\mathbf{Ctr_+(\sigma)}$. In the sequel $\delta = \langle v_1, \dots, v_k\rangle \in \Delta^\sigma$ is a circuit. Let $\Gamma \subset \Gamma_\sigma = K^*$ be a finite group. Denote by π (resp. π_Γ) the projection corresponding to the quotient $X_\delta \to X_\delta /\!/ K^*$ (resp. $X_\delta \to X_\delta/\Gamma$). Write $\pi(\delta) = \langle w_1, \dots, w_k\rangle$ and let $\sum_{r_i'>0} r_i' w_i = 0$ be the unique relation between vectors $(**)$ as in Section 4.4. Set $\mathrm{Ctr}_+(\delta) = \sum_{r_i'>0} w_i \in \overline{\mathrm{par}}(\pi(\delta_+)) \cap \mathrm{int}(\pi(\delta_+))$, where $\delta_+ = \langle v_i \mid r_i > 0\rangle$.

Denote by $\widehat{X}_\delta$ the completion of $\widetilde{X}_{\Delta^\sigma}$ at O_δ. By Corollary 5.36, the generic point $O_\delta \in \widetilde{X}_{\Delta^\sigma}$ is G_σ-invariant and thus G_σ acts on $\widehat{X}_\delta$. Moreover $K[\widehat{X}_\delta] = K(O_\delta)[[\delta^\vee]]$ is faithfully flat over a $\mathcal{O}_{\widetilde{X}_{\Delta^\sigma}, O_\delta}$. Also, $\widehat{\mathcal{O}}_{X_{\pi_\Gamma(\Delta)}, \widetilde{O}_{\pi_\Gamma(\delta)}} = K(\widetilde{O}_{\pi_\Gamma(\delta)})[[\underline{\pi_\Gamma(\delta)}^\vee]]$ is faithfully flat over $\mathcal{O}_{X_{\pi_\Gamma(\Delta)}, \widetilde{O}_{\pi_\Gamma(\delta)}}$ and we get

Lemma 5.37. *The valuation* $\mathrm{val}(v)$*, where* $v \in \pi_\Gamma(\delta)$*, is* G_σ*-invariant on* $\widehat{X}_\delta/\Gamma$ *iff it is* G_σ*-invariant on* $\widetilde{X}_{\Delta^\sigma}/\Gamma$.

Lemma 5.38. $\overline{O}_{\delta_-}, \overline{O}_{\delta_+} \subset \widehat{X}_\delta$ *and* $\overline{O}_{\delta_-}, \overline{O}_{\delta_+} \subset \widetilde{X}_{\Delta^\sigma}$ *are* G_σ*-invariant.*

Proof. By Lemmas 4.13 and 2.10, the ideal $I_{\overline{O}_{\delta_+}} \subset K[\widehat{X}_\sigma]$ of $\overline{O}_{\delta_+} = (O_\delta)^+$ is generated by functions with positive weights. □

Proposition 4.12, Lemma 4.13 and the above imply:

Corollary 5.39. *The morphisms* $\widehat{\phi}_-\colon (\widehat{X}_\delta)_-/K^* \to \widehat{X}_\delta /\!/ K^*$ *and* $\widehat{\phi}_+\colon (\widehat{X}_\delta)_+/K^* \to \widehat{X}_\delta /\!/ K^*$ *are* G_σ*-equivariant, proper and birational.*

Lemma 5.40. *The vector* $v := \mathrm{Mid}\,(\mathrm{Ctr}_+(\delta), \delta) = \pi^{-1}_{|\partial_-(\delta)}(\mathrm{Ctr}_+(\delta)) + \pi^{-1}_{|\partial_+(\delta)}(\mathrm{Ctr}_+(\delta))$ *is stable.*

Proof. Set $v_- := \pi^{-1}_{|\partial_-(\delta)}(\mathrm{Ctr}_+(\delta))$ and $v_+ := \pi^{-1}_{|\partial_+(\delta)}(\mathrm{Ctr}_+(\delta))$. By Lemma 5.38, $\overline{O}_{\delta_+} \subset \widetilde{X}_{\Delta^\sigma}$ is G_σ-invariant and, by Corollary 5.35(2) and Lemma 5.37, $\mathrm{val}(v_+)$ is G_σ-invariant on $\widetilde{X}_\sigma$ and on $\widehat{X}_\delta$. Hence the valuation $\mathrm{val}(v_+)$ descends to a G_σ-invariant valuation $\mathrm{val}(\pi(v_+))$ on $\widehat{X}_\delta /\!/ K^* = K(O_\delta)[[\delta^\vee]]^{K^*}$. By Corollary 5.39, $\mathrm{val}(\pi(v_-)) = \mathrm{val}(\pi(v_+))$ is G_σ-invariant on $(\widehat{X}_\delta)_+/K^* = \widehat{X}_{\partial_-(\delta)}/K^* = \widehat{X}_{\pi(\partial_-(\delta))}$. Let $\Gamma \subset K^*$ be the subgroup generated by all subgroups $\Gamma_\tau \subset K^*$, where $\tau \in \partial_-(\delta)$. Then K^*/Γ acts freely on $X_{\partial_-(\delta)}/\Gamma = (X_\delta)_+/\Gamma$. Let $j\colon (X_\delta)_+/\Gamma \to (X_\delta)_+/K^*$ be the natural morphism. Let $\pi_\Gamma\colon \delta \to \pi_\Gamma(\delta)$ be the projection corresponding to the quotient $X_\delta \to X_\delta/\Gamma$. By Lemma 5.7, for any $\tau \in \partial_-(\sigma)$, the restriction of j to $X_\tau/\Gamma \subset (X_\delta)_+/\Gamma$ is given by $j\colon X_\tau/\Gamma = X_{\underline{\tau}}/\Gamma \times O_\tau/\Gamma \to X_\tau/K^* = X_{\underline{\tau}}/\Gamma \times O_\tau/K^*$. Thus $\mathcal{I}_{\mathrm{val}(\pi_\Gamma(v_-)),a} = \hat{j}^*(\mathcal{I}_{\mathrm{val}(\pi(v_-)),a})$, where $\hat{j}\colon (\hat{X}_\delta)_+/\Gamma \to (\hat{X}_\delta)_+/K^*$ is the natural morphism induced by j. Since the morphism $\hat{j}$ is G_σ-equivariant it follows that $\mathrm{val}(\pi_\Gamma(v_-))$ is G_σ-equivariant on $(\widehat{X}_\delta)_+/\Gamma$. Since $(\widehat{X}_\delta)_+ \subset \widehat{X}_\delta$ is an open G_σ-equivariant inclusion and Γ is finite we get that the morphism $(\widehat{X}_\delta)_+/\Gamma \subset (\widehat{X}_\delta)/\Gamma$ is an open G_σ-equivariant inclusion. Thus the valuation $\mathrm{val}(\pi(v_-))$ is G_σ-equivariant on $\widehat{X}_\delta/\Gamma$ and on $\widetilde{X}_{\Delta^\sigma}/\Gamma$ (Lemma 5.37). Finally, by Lemma 5.29, $\mathrm{val}(v_-)$ it is G_σ-equivariant on $\widetilde{X}_{\Delta^\sigma}$. Thus by the convexity $\mathrm{val}(v) = \mathrm{val}(v_+ + v_-)$ is G_σ-equivariant on $\widetilde{X}_{\Delta^\sigma}$. □

5.11. π-desingularization of cobordisms. Let $\delta_1, \dots, \delta_k \in \Sigma$ be the circuits in Σ. Note that common faces of distinct circuits are independent. Also, every independent $\tau \in \Sigma$ is a face of some circuit $\tau < \delta$. Thus π-desingularization of circuits δ_i will determine π-desingularization of all faces in Σ. Apply Morelli π-desingularization to δ_1 to get $\Delta_1^{\sigma_1} := \langle v_{r_1}\rangle \dots \langle v_1\rangle \cdot \delta_1$. This defines a canonical subdivision Δ_1 of Σ, where $\Delta_1 := \langle v_{r_1}\rangle \dots \langle v_1\rangle \cdot \Sigma$. Next apply the π-desingularization to the subdivision $\Delta_1^{\sigma_2}$ of σ_2 to get $\Delta_2^{\sigma_2} := \langle v_{r_2}\rangle \dots \langle v_{r_1+1}\rangle \cdot \Delta_1^{\sigma_2}$ and $\Delta_2 = \langle v_{r_2}\rangle \dots \langle v_{r_1+1}\rangle \cdot \Delta_1$. Continue the process for other circuits to get the π-nonsingular subdivision $\Delta_k = \langle v_{r_k}\rangle \dots \langle v_1\rangle \cdot \Sigma$ of Σ.

5.12. Proof of the Weak Factorization Theorem. The decomposition Δ of Σ is obtained by a sequence of star subdivisions at stable centers (Lemmas 5.40, 5.34). By Propositions 5.23 and 5.18, Δ defines a birational projective modification $f\colon B^\pi \to B$. The modification does not affect points with trivial stabilizers $B_- = X^- \setminus X$ and $Y^+ \setminus Y$ (see Proposition 2.12). This means that $(B^\pi)_- = B_-$ and $(B^\pi)_+ = B_+$ and B^π is a cobordism between X and Y. Moreover B^π admits a projective compactification $\overline{B^\pi} = B^\pi \cup X \cup Y$. The cobordism $B^\pi \subset \overline{B^\pi}$ admits a decomposition into elementary cobordisms B_a^π, defined by the strictly increasing function χ_{B^π}. Let $F \in \mathcal{C}((B_a^\pi)^{K^*})$ be a fixed point component and $x \in F$ be a point. By Proposition 5.18 the modification $f\colon B^\pi \to B$ is locally described for a toric chart $\phi_\sigma\colon U \to X_\sigma$ by a smooth Γ_σ-equivariant morphism $\phi_{\Delta^\sigma}\colon f^{-1}(U) \to X_{\Delta^\sigma}$. Then by Lemma 4.4, $\phi_{\Delta^\sigma}(x)$ is in O_δ, where $\delta \in \Delta^\sigma$ is dependent and π-nonsingular. In particular the cone $\sigma \in \Sigma$ is also dependent and $\Gamma_\sigma = K^*$. So we locally have a smooth K^*-equivariant morphism

$$\phi_\delta\colon V_x \to X_\delta,$$

where $V_x \subset \phi_{\Delta^\sigma}^{-1}(X_\delta)$ is an affine K^*-invariant subset of B_a^π. This gives a diagram

$$\begin{array}{ccccc}
(B_a^\pi)_-/K^* & \supset & V_{x-}/K^* & \to & X_{\delta-}/K^* \\
\uparrow{\scriptstyle \psi_-} & & \uparrow & & \uparrow{\scriptstyle \phi_-} \\
\Gamma((B_a^\pi)_-/K^*, (B_a^\pi)_+/K^*) & \supset & \Gamma(V_{x-}/K^*, V_{x+}/K^*) & \to & \Gamma(X_{\delta-}/K^*, X_{\delta+}/K^*) \\
\downarrow{\scriptstyle \psi_+} & & \downarrow & & \downarrow{\scriptstyle \phi_+} \\
(B_a^\pi)_+/K^* & \supset & V_{x+}/K^* & \to & X_{\delta+}/K^*
\end{array}$$

with horizontal arrows smooth. Here $\Gamma(X_-/K^*, X_+/K^*)$ denotes the normalization of the graph of a birational map $X_-/K^* \dashrightarrow X_+/K^*$ for a relevant cobordism X. We use functoriality of the graph (a dominated component of the fiber product $X_-/K^* \times_{X//K^*} X_+/K^*$). By Corollary 4.16 the morphisms ϕ_- and ϕ_+ are blow-ups at smooth centers. Thus ψ_- and ψ_+ are locally blow-ups at smooth centers so they are globally blow-ups at smooth centers.

References

[1] Abhyankar, S., On the valuations centered in a local domain. *Amer. J. Math.* **78** (1956), 321–348.

[2] Abramovich, D., and de Jong, A. J., Smoothness, semistability, and toroidal geometry. *J. Algebraic Geom.* **6** (1997), 789–801.

[3] Abramovich, D., Karu, K., K. Matsuki, J. Włodarczyk, Torification and factorization of birational maps. *J. Amer. Math. Soc.* **15** (2002), 531–572.

[4] Abramovich, D., Matsuki, K., and Rashid, S., A note on the factorization theorem of toric birational maps after Morelli and its toroidal extension. *Tohoku Math. J.* (2) **51** (1999), 489–537.

[5] Białynicki-Birula, A., Święcicka, J., Complete quotients by algebraic torus actions. In *Group actions and vector fields*, Lecture Notes in Math. 956, Springer-Verlag, Berlin 1982, 10–21.

[6] Bierstone, E., and Milman, D., Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant. *Invent. Math.* **128** (1997), 207–302.

[7] M. Brion and C. Procesi, Action d'un tore dans une variété projective. in *Operator algebras, unitary representations, enveloping algebras, and invariant theory* (Paris, 1989), Progr. Math. 92, Birkhäuser, Boston 1990, 509–539.

[8] Christensen, C., Strong domination/weak factorization of three dimensional regular local rings. *J. Indian Math. Soc.* **45** (1981), 21–47.

[9] Cutkosky, S. D., Local factorization of birational maps. *Adv. in Math.* **132** (1997), 167–315.

[10] Cutkosky, S. D., Local factorization and monomialization of morphisms. *Astérisque* **260** (1999).

[11] Danilov, V. I., The geometry of toric varieties. *Russian Math. Surveys* **33** (1978), 97–154.

[12] Danilov, V. I., Birational geometry of toric 3-folds. *Math. USSR-Izv.* **21** (1983), 269–280.

[13] Dolgachev, I. V., and Hu, Y., Variation of geometric invariant theory quotients. *Inst. Hautes Études Sci. Publ.* Math. **87** (1998), 5–56.

[14] Ewald, G., Blow-ups of smooth toric 3-varieties. *Abh. Math. Sem. Univ. Hamburg* **57** (1987), 193–201.

[15] Fulton, W., *Introduction to Toric Varieties*. Ann. of Math. Stud. 131, Princeton University Press, Princeton 1993.

[16] Hartshorne, R., *Algebraic Geometry*. Grad. Texts in Math. 52, Springer-Verlag, New York, Heidelberg 1977.

[17] Hironaka, H., On the theory of birational blowing-up. Harvard University Ph.D. Thesis, 1960.

[18] Hironaka, H., Resolution of singularities of an algebraic variety over a field of characteristic zero. *Ann. of Math.* **79** (1964), 109–326.

[19] Hu, Y., The geometry and topology of quotient varieties of torus actions. *Duke Math. J.* **68** (1992), 151–184; Erratum ibid. **68** (1992), 609.

[20] Hu, Y., Relative geometric invariant theory and universal moduli spaces. *Internat. J. Math.* **7** (1996), 151–181.

[21] Hu, Y., and Keel, Ś., A GIT proof of Włodarczyk's weighted factorization theorem. Preprint; math.AG/9904146.

[22] Karu, K., Local strong factorization of toric birational maps. *J. Algebraic Geom.* **14** (2005), 165–175.

[23] Kempf, G., Knudsen, F., Mumford, D., and Saint-Donat, B., *Toroidal embeddings I.* Lecture Notes in Math. 339, Springer-Verlag, Berlin 1973.

[24] Luna, D., Slices étales. *Sur les groupes algébriques*, *Bull. Soc. Math. France* **101** (33) (1973), 81–105.

[25] Matsuki, K., Lectures on factorization of birational maps. *Publ. Res. Inst. Math. Sci.*, preprint, 1999.

[26] Milnor, J., *Morse Theory*. Ann. of Math. Stud. 51, Princeton University Press, Princeton 1963.

[27] Morelli, R., The birational geometry of toric varieties. *J. Algebraic Geom.* **5** (1996), 751–782.

[28] Morelli, R., Correction to "The birational geometry of toric varieties". 2000; http://www.math.utah.edu/~morelli/Math/math.html.

[29] Mumford, D., Fogarty, J., and Kirwan, F., *Geometric Invariant Theory*. Third enlarged edition, Ergeb. Math. Grenzgeb. 34, Springer-Verlag, Berlin 1994.

[30] Oda, T., *Torus embeddings and applications*. Based on joint work with Katsuya Miyake, Tata Inst. Fund. Res., Bombay, 1978.

[31] Oda, T., *Convex Bodies and Algebraic Geometry*. Ergeb. Math. Grenzgeb. 15, Springer-Verlag, Berlin 1988.

[32] Reid, M., Decomposition of Toric Morphisms. In *Arithmetic and Geometry* (ed. by M. Artin and J. Tate), Progr. Math. 36, Birkhäuser, Boston 1983, 395–418.

[33] Sumihiro, H., Equivariant Completion I, II. *J. Math. Kyoto Univ.* **14** (1974), 1–28; **15** (1975) 573–605.

[34] Thaddeus, M., Geometric invariant theory and flips. *J. Amer. Math. Soc.* **9** (1996), 691–723.

[35] Villamayor, O., Constructiveness of Hironaka's resolution. *Ann. Sci. École Norm. Sup.* (4) **22** (1) (1989), 1–32.

[36] Włodarczyk, J., Decomposition of birational toric maps in blow-ups and blow-downs. A proof of the Weak Oda Conjecture. *Trans. Amer. Math. Soc.* **349** (1997), 373–411.

[37] Włodarczyk, J., Birational cobordisms and factorization of birational maps. *J. Algebraic Geom.* **9** (2000), 425–449.

[38] Włodarczyk, J., Toroidal varieties and the weak factorization theorem. *Invent. Math.* 154 (2003), 223–231.

[39] Włodarczyk, J., Simple constructive weak factorization. Preprint; AG/0601649.

[40] Włodarczyk, J., Simple Hironaka resolution in characteristic zero. *J. Amer. Math. Soc.* **18** (2005), 779–822.

[41] Zariski, O., *Algebraic Surfaces*. Ergeb. Math. Grenzgeb. 3, Springer-Verlag, Berlin 1935.

Department of Mathematics, Purdue University, West Lafayette, IN 47907, U.S.A.
E-mail: wlodar@math.purdue.edu

Manifolds with positive curvature operators are space forms

Christoph Böhm and Burkhard Wilking

Abstract. We show that the normalized Ricci flow evolves metrics with positive curvature operator on compact manifolds to limit metrics of constant positive sectional curvature. In this note we only indicate the proof, the details will be published somewhere else.

Mathematics Subject Classification (2000). Primary 53C44; Secondary 53C25.

Keywords. Ricci flow, positive curvature operator, space forms.

1. Introduction

The Ricci flow has been introduced by Hamilton in 1982 [H1] in order to prove that a compact 3-manifold admitting a Riemannian metric of positive Ricci curvature is a spherical space form. The unnormalized Ricci flow is the geometric evolution equation

$$\frac{\partial g}{\partial t} = -2\,\mathrm{Ric}(g) \tag{1}$$

for a curve g_t of Riemannian metrics on a compact manifold M^n.

Using moving orthonormal frames, this leads to the following evolution equation for the curvature operator $\mathrm{R}_t\colon \Lambda^2 T_pM \to \Lambda^2 T_pM \cong \mathfrak{so}(T_pM)$ of g_t (cf. [H2]):

$$\frac{\partial\,\mathrm{R}}{\partial t} = \Delta\,\mathrm{R} + 2(\mathrm{R}^2 + \mathrm{R}^{\#}), \tag{2}$$

where $\mathrm{R}^{\#} = \mathrm{ad}\circ(\mathrm{R}\wedge\mathrm{R})\circ\mathrm{ad}^*$ and $\mathrm{ad}\colon \Lambda^2(\mathfrak{so}(T_pM)) \to \mathfrak{so}(T_pM)$ is the adjoint representation. For the details we refer the reader to Section 2. By Hamilton's maximum principle certain dynamical properties of the partial differential equation (2) can be derived from the dynamical properties of the corresponding ordinary differential equation

$$\frac{d\,\mathrm{R}}{dt} = \mathrm{R}^2 + \mathrm{R}^{\#}. \tag{3}$$

In dimension four Hamilton showed that compact 4-manifolds with positive curvature operators are spherical space forms as well [H2]. More generally, Chen showed that the same conclusion holds for compact 4-manifolds with 2-positive curvature

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

operator [Che]. Recall that a curvature operator is called 2-positive, if the sum of its two smallest eigenvalues is positive. In arbitrary dimensions it was shown by Huisken [Hu], that there is an explicit open cone in the space of curvature operators such that the normalized Ricci flow evolves metrics whose curvature operator at each point is contained in that cone into metrics of constant positive sectional curvature.

All of these results are also based on the maximum principle. The main reason why they are more involved is that the ordinary differential equation (3) is not that well understood in dimensions $n > 3$ and in particular not for $n > 4$.

Hamilton conjectured that in all dimensions compact Riemannian manifolds with positive curvature operators must be space forms. We can confirm Hamilton's conjecture. More generally, we can show the following

Theorem 1. *Let (M, g) be a compact Riemannian manifold with 2-positive curvature operator. Then the normalized Ricci flow evolves g to a limit metric with constant positive sectional curvature.*

The theorem is known in dimensions below five and our proof only works in dimensions above two. Since the proof solely relies on Hamilton's maximum principle it carries over to orbifolds.

Let us mention that this is no longer true in dimension two. By the work of Hamilton [H4] and Chow [Cho] it is known that the normalized Ricci flow converges to a metric of constant curvature for any initial metric in the manifold case. However, there exist two dimensional orbifolds with positive sectional curvature which are not covered by a manifold. On such orbifolds the Ricci flow converges to a nontrivial Ricci soliton [CW].

There is a wealth of different techniques in geometry to prove sphere theorems. Here we only mention the theorem of Micallef and Moore [MM] that a simply connected manifold with positive isotropic curvature is a homotopy sphere. It is well known that a 2-positive curvature operator has positive isotropic curvature. However, the techniques of Micallef and Moore do not allow to get restrictions for the fundamental groups or the differentiable structure.

Let us turn to the proof of Theorem 1. The major obstacle is to understand the ordinary differential equation (3). Here we establish a new algebraic identity which should be useful in other context as well. We study how the differential equation changes if we pull it back by an equivariant linear map $1\colon S^2_B(\mathfrak{so}(n)) \to S^2_B(\mathfrak{so}(n))$, where $S^2_B(\mathfrak{so}(n))$ denotes the space of curvature operators satisfying the Bianchi identity. For $n \geq 4$ this space decomposes into three pairwise inequivalent $\mathrm{O}(n)$-invariant irreducible subspaces

$$S^2_B(\mathfrak{so}(n)) = \langle I \rangle \oplus \langle \mathrm{Ric}_0 \rangle \oplus \langle \mathrm{W} \rangle.$$

Here $\langle I \rangle$ denotes multiples of the identity, $\langle \mathrm{W} \rangle$ curvature operators with vanishing Ricci curvature and $\langle \mathrm{Ric}_0 \rangle$ are the curvature operators of traceless Ricci type. Given such a curvature operator R we let R_I, $\mathrm{R}_{\mathrm{Ric}_0}$ and R_{W}, denote the projections onto $\langle I \rangle$, $\langle \mathrm{Ric}_0 \rangle$ and $\langle \mathrm{W} \rangle$, respectively.

Since we can always scale the linear map l by a factor it is not much of a restriction to consider maps of the form

$$\mathrm{l}_{a,b}(\mathrm{R}) = \mathrm{R} + 2(n-1)a\,\mathrm{R}_I + (n-2)b\,\mathrm{R}_{\mathrm{Ric}_0}. \tag{4}$$

Notice that $\mathrm{l}_{a,b}$ induces the identity on the space $\langle \mathrm{W} \rangle$ of Weyl curvature operators.

The following result is crucial.

Theorem 2. *Let* $\mathrm{R} \in S^2_B(\mathfrak{so}(n))$ *be a curvature operator,* $\mathrm{Ric} \in S^2\mathbb{R}^n$ *its Ricci curvature,* Ric_0 *the traceless part of* Ric*, and let*

$$D_{a,b} = \mathrm{l}_{a,b}^{-1}\big((\mathrm{l}_{a,b}\,\mathrm{R})^2 + (\mathrm{l}_{a,b}\,\mathrm{R})^{\#}\big) - \mathrm{R}^2 - \mathrm{R}^{\#}.$$

Then

$$\begin{aligned} D_{a,b} = {} & \big((n-2)b^2 - 2(a-b)\big)\,\mathrm{Ric}_0 \wedge \mathrm{Ric}_0 + 2a\,\mathrm{Ric} \wedge \mathrm{Ric} + 2b^2\,\mathrm{Ric}_0^2 \wedge \mathrm{id} \\ & + \frac{\mathrm{tr}(\mathrm{Ric}_0^2)}{n + 2n(n-1)a}\big(nb^2(1-2b) - 2(a-b)(1-2b+nb^2)\big)\,I. \end{aligned}$$

In particular, $D_{a,b}$ *is independent of the Weyl curvature of* R.

The theorem often allows to construct new invariant curvature conditions by considering the image of a known invariant curvature conditions under the linear map $\mathrm{l}_{a,b}$ for suitable constants $a, b \in \mathbb{R}$. Recall that an invariant curvature condition is a convex subset of $S_B(\mathfrak{so}(n))$ which is invariant under the ordinary differential equation (3) and hence, by Hamilton's maximum principle, invariant under the partial differential equation (2).

As already mentioned this note only contains indications of proofs. The details will be published somewhere else. We expect that the new algebraic identity on curvature operators (Theorem 2) and its Kähler analogue should give rise to further applications. This will be the subject of a forthcoming paper.

2. Preliminaries

For a Euclidean vector space V we let $\Lambda^2 V$ denote the exterior product of V. We endow $\Lambda^2 V$ with its natural scalar product; if $e_1, \ldots, e_n$ is an orthonormal basis of V then $e_1 \wedge e_2, \ldots, e_{n-1} \wedge e_n$ is an orthonormal basis of $\Lambda^2 V$. Notice that two linear endomorphisms A, B of V induce a linear map

$$A \wedge B\colon \Lambda^2 V \to \Lambda^2 V, \quad v \wedge w \mapsto \tfrac{1}{2}\big(A(v) \wedge B(w) + B(v) \wedge A(w)\big).$$

We will identify $\Lambda^2\mathbb{R}^n$ with the Lie algebra $\mathfrak{so}(n)$ by mapping the unit vector $e_i \wedge e_j$ onto the linear map $L(e_i \wedge e_j)$ of rank two which is a rotation with angle $\pi/2$ in the plane spanned by e_i and e_j. Notice that under this identification the scalar product on $\mathfrak{so}(n)$ corresponds to $\langle A, B \rangle = -1/2\,\mathrm{tr}(AB)$.

We let $S^2(\mathfrak{so}(n))$ denote the space of selfadjoint endomorphisms of $\mathfrak{so}(n)$ and $S^2_B(\mathfrak{so}(n))$ the subspace of operators satisfying the Bianchi identity. Recall that $S^2_B(\mathfrak{so}(n))$ is given by the orthogonal complement of $\Lambda^4\mathbb{R}^n$ in $S^2(\mathfrak{so}(n))$. We use the convention that the curvature operator of the round sphere is the identity of $\mathfrak{so}(n)$. This explains the extra factor 2 in the evolution equation (2), cf. [H2].

Hamilton observed that for two elements $\mathrm{S}, \mathrm{R} \in S^2(\mathfrak{so}(n))$ one can define a new element $\mathrm{S}\,\#\,\mathrm{R} \in S^2(\mathfrak{so}(n))$. This can be done invariantly by putting

$$\mathrm{S}\,\#\,\mathrm{R} := \mathrm{ad} \circ (\mathrm{S} \wedge \mathrm{R}) \circ \mathrm{ad}^*$$

where $\mathrm{ad}\colon \Lambda^2\mathfrak{so}(n) \to \mathfrak{so}(n),\ X \wedge Y \mapsto [X, Y]$ is the adjoint representation and ad^* is its dual. Following Hamilton we use the abbreviation $\mathrm{R}^{\#} = \mathrm{R}\,\#\,\mathrm{R}$. For a generic $\mathrm{R} \in S^2_B(\mathfrak{so}(n))$ neither R^2 nor $\mathrm{R}^{\#}$ are in $S^2_B(\mathfrak{so}(n))$. However, Hamilton showed that the sum is in $S^2_B(\mathfrak{so}(n))$ and hence the ordinary differential equation (3) leaves $S^2_B(\mathfrak{so}(n))$ invariant.

Recall that an $\mathrm{O}(n)$-invariant subset C in the space of curvature operators is invariant under the Ricci flow, if the Ricci flow evolves metrics on compact manifolds whose curvature operator lies in C at any point into metrics with the same property.

Theorem 2.1 (Hamilton, [H2]). *A closed convex $\mathrm{O}(n)$-invariant subset $C \subset S_B(\mathfrak{so}(n))$ of curvature operators is invariant under the Ricci flow if it is invariant under the ordinary differential equation*

$$\frac{d\,\mathrm{R}}{dt} = \mathrm{R}^2 + \mathrm{R}^{\#}.$$

We recall that if $e_1, \ldots, e_n$ denotes an orthonormal basis, then

$$\mathrm{Ric}(\mathrm{R}^2 + \mathrm{R}^{\#})_{ij} = \sum_{k,l} \mathrm{Ric}_{kl}\, \mathrm{R}_{kijl} \tag{5}$$

where $\mathrm{R}_{kijl} = \langle \mathrm{R}(e_i \wedge e_k), e_j \wedge e_l \rangle$, see [H1], [H2].

Finally let us mention that we learned from Huisken that the trilinear map

$$\mathrm{tri}(\mathrm{R}_1, \mathrm{R}_2, \mathrm{R}_3) := \mathrm{tr}\big((\mathrm{R}_1\,\mathrm{R}_2 + \mathrm{R}_2\,\mathrm{R}_1 + 2\,\mathrm{R}_1\,\#\,\mathrm{R}_2)\,\mathrm{R}_3\big) \tag{6}$$

is symmetric in all three components $\mathrm{R}_1, \mathrm{R}_2, \mathrm{R}_3 \in S^2(\mathfrak{so}(n))$. Moreover (3) corresponds to the gradient flow of the function $\frac{1}{6}\,\mathrm{tri}(\mathrm{R}, \mathrm{R}, \mathrm{R})$.

3. On the proof of Theorem 2

In this section we show that the difference tensor $D = D_{a,b}$ does not depend on the Weyl curvature of R. The precise formula in Theorem 2 then follows from a calculation. We view D as quadratic form in R. By

$$B(\mathrm{R}, \mathrm{S}) := \tfrac{1}{4}\big(D(\mathrm{R} + \mathrm{S}) - D(\mathrm{R} - \mathrm{S})\big)$$

we get the corresponding bilinear form.

Let $S = W \in \langle W \rangle$. We have to show $B(R, W) = 0$ for all $R \in S^2_B(\mathfrak{so}(n))$. We start by considering $R \in \langle W \rangle$. Then $l_{a,b}(R \pm W) = R \pm W$. It follows from the formula (5) for the Ricci curvature of $R^2 + R^\#$ that $(R \pm W)^2 + (R \pm W)^\#$ has vanishing Ricci tensor. Hence $(R \pm W)^2 + (R \pm W)^\#$ is a Weyl curvature operator and accordingly fixed by $l^{-1}_{a,b}$. Therefore $B(R, W) = 0$.

Next we consider the case that $R = I$ is the identity. Notice that

$$(I + W)^2 + (I + W)^\# - (I - W)^2 - (I - W)^\# = 4W + 4W \# I = 0.$$

The last equation follows by a straightforward computation for $n = 4$. Since there is a natural embedding of the Weyl tensors of $S^2_B(\mathbb{R}^4)$ to the Weyl tensors of $S^2_B(\mathbb{R}^n)$ the same holds for $n \geq 5$. Clearly the equation implies $B(W, I) = 0$.

It remains to consider the case of $R \in \langle \mathrm{Ric}_0 \rangle$. Using the symmetry of the trilinear form (6) we see for each $W_2 \in \langle W \rangle$ that

$$\mathrm{tri}(W, R, W_2) = \mathrm{tri}(W, W_2, R) = 0$$

because $W W_2 + W_2 W + 2 W \# W_2$ lies in $\langle W \rangle$ and $R \in \langle \mathrm{Ric}_0 \rangle$. Combining this with $\mathrm{tri}(W, R, I) = 0$ gives that $WR + RW + 2W \# R \in \langle \mathrm{Ric}_0 \rangle$. Using once more that $l := l_{a,b}$ is the identity on $\langle W \rangle$ we see that

$$l(W)\, l(R) + l(R)\, l(W) + 2\, l(W) \# l(R) = l(WR + RW + 2W \# R).$$

This clearly proves $B(W, R) = 0$.

4. On the proof of Theorem 1

We call a continuous family $C(t)_{t \in [0,1)} \subset S^2_B(\mathfrak{so}(n))$ of closed convex $\mathrm{O}(n)$-invariant cones of full dimension a pinching family, if

1. each $R \in C(t) \setminus \{0\}$ has positive scalar curvature,
2. $R^2 + R^\#$ is contained in the interior of the tangent cone of $C(t)$ at R for all $R \in C(t) \setminus \{0\}$ and all $t \in (0, 1)$,
3. $C(t)$ converges in the pointed Hausdorff topology to the one-dimensional cone $\mathbb{R}^+ I$ as $t \to 1$.

Using Theorem 2 we can show that

Theorem 4.1. *There is a pinching family $C(t)_{t \in [0,1)} \subset S^2_B(\mathfrak{so}(n))$ such that $C(0)$ is given by the cone of 2 nonnegative curvature operators.*

Outline of the proof of Theorem 4.1. Here we only show that we can stay away from the boundary of the cone C of 2 nonnegative curvature operators.

$$b \in \left(0, \tfrac{\sqrt{2n(n-2)+4}-2}{n(n-2)}\right] \quad \text{and} \quad 2a = 2b + (n-2)b^2$$

We claim that then the set $\mathrm{l}_{a,b}(C)$ is invariant under the ordinary differential equation. We have to show that for $\mathrm{R} \in C$ the curvature operator

$$X_{a,b} = \mathrm{l}_{a,b}^{-1}(\mathrm{l}_{a,b}(\mathrm{R})^2 + \mathrm{l}_{a,b}(\mathrm{R})^{\#})$$

is contained in the tangent cone $T_{\mathrm{R}}C$ of C at R. Notice that by assumption we have $\mathrm{R}^2 + \mathrm{R}^{\#} \in T_{\mathrm{R}}C$. Thus it suffices to show that $D_{a,b} = X_{a,b} - \mathrm{R}^2 - \mathrm{R}^{\#}$ lies in $T_{\mathrm{R}}C$. For that it is clearly sufficient to show that $D_{a,b}$ is positive definite. We know that $\mathrm{Ric}(\mathrm{R}) \geq 0$ for any $\mathrm{R} \in C$. Looking at the formula for $D_{a,b}$ in Theorem 2 it suffices to show that

$$0 \leq b^2\big(n(1-2b) - (n-2)(1-2b+nb^2)\big)$$

holds in the given range. This is a straightforward computation.

It is not hard to see that for $b = \frac{\sqrt{2n(n-2)+4}-2}{n(n-2)}$ and $n \geq 4$ the closed cone $\mathrm{l}_{a,b}(C)$ is contained in the open cone of positive curvature operators. Thus it suffices to prove the existence of a pinching family of cones with $C(0)$ being the cone of nonnegative curvature operators. Theorem 2 combined with a calculation, which will be carried out somewhere else, shows that such a family can be constructed in the form

$$C(t) := \mathrm{l}_{a(t),b(t)}\big(\big\{\mathrm{R} \mid \mathrm{R} \geq 0, \mathrm{Ric}(\mathrm{R}) \geq p(t)\tfrac{\mathrm{scal}(\mathrm{R})}{n}\big\}\big)$$

for suitable functions $a(t)$, $b(t)$ and $p(t)$. □

Theorem 1 then follows from Theorem 4.1 combined with

Theorem 4.2. *Let $C(t)_{t\in[0,1)} \subset S^2_B(\mathfrak{so}(n))$ be a pinching family of closed convex cones, $n \geq 3$. Suppose that (M, g) is a compact Riemannian manifold such that the curvature operator of M at each point is contained in the interior of $C(0)$. Then the normalized Ricci flow evolves g to a constant curvature limit metric.*

On the proof of Theorem 4.2. Since M is compact and the family of cones is continuous we can assume that for a sufficiently large h_0 and a sufficiently small ε we have for the curvature operator R_p of (M, g) at each point $p \in M$ that

$$\mathrm{R}_p \in \{\mathrm{R} \mid \mathrm{scal} \leq h_0\} \cap C(\varepsilon).$$

We now consider to h_0 and ε the intersection F of all convex subsets which contain the above set and which are invariant under the ordinary differential equation (3). Clearly F is then invariant under the ordinary differential equation and one can show that for each t the set $F \setminus C(t)$ is relatively compact.

Thus F is a generalized pinching set similarly to Hamilton's concept.

From the maximum principle we know that the Ricci flow evolves g to metrics g_t whose curvature operators at each point are contained in F. We do also know that the solution of the Ricci flow exists as long as the curvature does not tend to infinity. Furthermore it follows from the maximum principle that the unnormalized Ricci flow exists only on a finite time interval $t \in [0, t_0)$. By Shi it follows from the maximum principle applied to the evolution equations for the i-th derivatives of the curvature tensor that

$$\max \|\nabla^i R_t\|^2 \leq C_i \max \|R_t\|^{i+2}$$

for all $t \in [t_0/2, t_0)$.

We now rescale each metric $g(t)$ to a metric $\tilde{g}(t)$ such that the maximal sectional curvature is equal to 1. From the above estimates it follows that we have a priori bounds for all derivatives of the curvature tensor of the metric $\tilde{g}(t)$ for $t \in [t_0/2, t_0)$.

We now look at a point $p_t \in (M, \tilde{g}_t)$ where the sectional curvature attains it maximum 1. We pull the metric via the exponential map back to the ball of radius π in $T_{p_t}M$. We identify this ball with the ball $B_\pi(0) \subset \mathbb{R}^n$ by choosing a linear isometry $\mathbb{R}^n \to T_{p_t}M$. We let $\bar{g}_t$ denote the induced metric on $B_\pi(0)$. From the above estimates on the derivatives of the curvature tensor it is clear that for any sequence t_n converging to t_0 there is a subsequence of $\bar{g}_{t_n}$ converging in the C^∞ topology.

The curvature operator at each point of the limit metric is contained in the set $\bigcap \frac{1}{\lambda_n^2} F = \mathbb{R}^+ I$, where λ_n denote the scaling factors which by assumption tend to infinity. Thus the limit metric on $B_\pi(0)$ has pointwise constant sectional curvature and by Schur's theorem it has constant curvature one, where we used $n \geq 3$.

It is now easy to deduce that the minimal sectional curvature of $(M, \tilde{g}_t)$ tends to 1 as well for $t \to t_0$. In particular for t close to t_0 we can apply Klingenberg's injectivity radius estimate for quarter pinched manifolds. Hence the universal cover of M has injectivity radius $\geq \pi$. This shows that the volume of the manifold does not converge to zero. It is then well known that we get a smooth limit space of constant curvature and that (M, g_t) is close in the C^∞-topology to this limit space. □

Finally we remark that the above proof carries over to orbifolds. In addition to Klingenberg's injectivity radius estimate one needs

Proposition 4.3. *Suppose (M, g) is compact orbifold with sectional curvature K. If $1/4 < K \leq 1$ and $\dim(M) \geq 3$ then M is the quotient of a Riemannian manifold by a finite isometric group action.*

The proof of the proposition is not related to the Ricci flow at all. It should rather be viewed as a generalization of Klingenberg's injectivity radius estimate.

References

[CE] Cheeger, J., Ebin, D., *Comparison theorems in Riemannian geometry*. North-Holland Mathematical Library 9, American Elsevier Publishing Co., New York 1975.

[Che] Chen, H., Pointwise 1/4-pinched 4-manifolds. *Ann. Global Geom.* **9** (1991), 161–176.

[Cho] Chow, B., The Ricci flow on the 2-sphere. *J. Differential Geom.* **33** (1991), 325–334.

[CW] Chow, B., Wu, L.-F., The Ricci flow on compact 2-orbifolds with curvature negative somewhere. *Comm. Pure. Appl. Math.* **44** (1991), 275–286.

[H1] Hamilton, R., Three-manifolds with positive Ricci curvature. *J. Differential Geom.* **17** (1982), 255–306.

[H2] Hamilton, R., Four-manifolds with positive curvature operator. *J. Differential Geom.* **24** (1986), 153–179.

[H3] Hamilton, R., The formation of singularities in the Ricci flow. In *Surveys in differential geometry* (Cambridge, MA, 1993), Vol. II, International Press, Cambridge, MA, 1995, 7–136.

[H4] Hamilton, R., The Ricci flow on surfaces. In *Mathematics and general relativity*, Contemp. Math. 71, Amer. Math. Soc., Providence, RI, 1988, 237–262.

[Hu] Huisken, G., Ricci deformation on the metric on a Riemannian manifold. *J. Differential Geom.* **21** (1985), 47–62.

[MM] Micallef, M., Moore, J. D., Minimal two-spheres and the topology of manifolds with positive curvature on totally isotropic two-planes. *Ann. of Math.* (2) **127** (1988), 199–227.

[Pe] Perelman, G., The entropy formula for the Ricci flow and its geometric applications. arXiv:math.DG/0307245.

[Shi] Shi, W., Deforming the metric on complete Riemannian manifolds. *J. Differential Geom.* **30** (1989), 223–301.

[Ta] Tachibana, S., A theorem on Riemannian manifolds of positive curvature operator. *Proc. Japan Acad.* **50** (1974), 301–302.

Westfälische Wilhelms-Universität, Einsteinstr. 62, 48149 Münster, Germany
E-mail: cboehm@math.uni-muenster.de

Westfälische Wilhelms-Universität, Einsteinstr. 62, 48149 Münster, Germany
E-mail: wilking@math.uni-muenster.de

Elliptic and parabolic problems in conformal geometry

Simon Brendle

Abstract. I will review recent results regarding two questions that arise in connection with the Yamabe problem. The first problem is concerned with the conformal deformation of Riemannian metrics by their scalar curvature. This leads to a parabolic evolution equation, which can be interpreted as the flow of steepest descent for the Yamabe functional. I will provide conditions which guarantee that the flow converges to a metric of constant scalar curvature as $t \to \infty$. The second problem is concerned with the set of constant scalar curvature metrics in a given conformal class. I will discuss under what conditions this set is compact. A recurring theme in the study of both problems is that blow-up can be ruled out by means of the positive mass theorem provided that the dimension is less than 6, whereas the Weyl tensor plays a crucial role in higher dimensions.

Mathematics Subject Classification (2000). Primary 53C21; Secondary 53C44.

Keywords. Scalar curvature; conformal deformation of Riemannian metrics; blow-up analysis.

1. The uniformization theorem and the Ricci flow in dimension 2

To begin with, I will discuss some results concerning the Ricci flow in dimension 2 and their relation to the uniformization theorem.

Theorem 1.1. *Let M be a compact manifold of dimension 2 without boundary. Given any metric g_0 on M, there exists a metric g which is pointwise conformal to g_0 and has constant curvature.*

Recall that two metrics g_0 and g on a manifold M are said to be pointwise conformal if there exists a smooth function $w\colon M \to \mathbb{R}$ such that $g = e^{2w}\, g_0$. If M is two-dimensional, then the Gaussian curvature of g is related to the Gaussian curvature of g_0 by the formula

$$K_g\, e^{2w} = K_{g_0} - \Delta_{g_0} w, \tag{1}$$

where Δ_{g_0} denotes the Laplacian relative to the metric g_0. Hence, the uniformization theorem in dimension 2 is equivalent to the solvability of the nonlinear elliptic equation

$$\Delta_{g_0} w - K_{g_0} + k\, e^{2w} = 0, \tag{2}$$

where k is a constant. Solutions to (2) can be constructed using variational methods (see e.g. [18], [19]). This approach is based on the observation that every solution

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of (2) is a critical point of the functional

$$E_{g_0}(w) = \frac{1}{2} \int_M |dw|^2_{g_0} \, d\text{vol}_{g_0} + \int_M K_{g_0} \, w \, d\text{vol}_{g_0} \\ - \pi \, \chi(M) \, \log \bigg(\int_M e^{2w} \, d\text{vol}_{g_0} \bigg), \tag{3}$$

where $\chi(M)$ denotes the Euler characteristic of M. The functional $E_{g_0}(w)$ was introduced by M. Berger in his work on the uniformization problem [5], and is known as the Liouville energy. It has a geometric interpretation in terms of the log determinant of the Laplacian associated with the conformal metric $e^{2w} \, g_0$ (see [27]).

The partial differential equation (2) is closely related to the Ricci flow in dimension 2. In dimension 2, the normalized Ricci flow takes the form

$$\frac{\partial}{\partial t} g(t) = -(K_{g(t)} - k_{g(t)}) \, g(t), \tag{4}$$

where $K_{g(t)}$ denotes the Gaussian curvature of $g(t)$ and

$$k_{g(t)} = \frac{2\pi \, \chi(M)}{\int_M d\text{vol}_{g(t)}}$$

is the mean value of the Gaussian curvature on M. The evolution equation (4) can be reduced to a nonlinear partial differential equation of parabolic type. Hence, given any initial metric g_0 on M, the evolution equation (4) has a smooth solution on a small time interval. The evolution equation (4) was first studied by R. Hamilton [14]. The longtime behavior of the flow is characterized by the following result due to R. Hamilton [14] and B. Chow [9]:

Theorem 1.2. *Let M be a compact manifold of dimension 2 without boundary. Given any initial metric g_0 on M, the evolution equation (4) has a global solution. The solution converges exponentially to a metric of constant curvature as $t \to \infty$.*

The convergence of the Ricci flow follows from the maximum principle if $\chi(M) \leq 0$. The case $\chi(M) > 0$ is more subtle. To prove Theorem 1.2 in this case, Hamilton assumed that the initial metric has positive curvature. This condition is preserved by the flow, and guarantees that the entropy

$$\int_M K_g \log K_g \, d\text{vol}_g$$

is well defined. It is shown in [14] that this functional is decreasing along the Ricci flow (see also [9]). In view of the Harnack inequality established in [14], this implies that the curvature is uniformly bounded from above. This is sufficient to prove Theorem 1.2 for initial metrics of positive curvature. The remaining cases were settled by B. Chow [9].

In a subsequent paper, Hamilton observed that the constant in the isoperimetric inequality improves along the Ricci flow [16]. This estimate can be used to give an alternative proof of Theorem 1.2. M. Struwe, building upon earlier work of X. Chen [8], provided another proof of Theorem 1.2 based on concentration-compactness arguments [39].

There are essentially two ways to generalize these results to higher dimensions. One is E. Calabi's problem concerning the existence of Kähler–Einstein metrics (or, more generally, extremal Kähler metrics). The other one is the Yamabe problem, which will be discussed below.

2. The Yamabe problem

In 1960, H. Yamabe [42] conjectured that the uniformization theorem can be generalized in the following way:

Theorem 2.1. *Let M be a compact manifold of dimension $n \geq 3$ without boundary, and let g_0 be a Riemannian metric on M. Then there exists a metric g which is pointwise conformal to g_0 and has constant scalar curvature.*

Theorem 2.1 was proved by T. Aubin [1], R. Schoen [29], and N. Trudinger [40]. A. Bahri gave an alternative proof of Theorem 2.1 in the locally conformally flat case [3].

As in the two-dimensional case, the Yamabe problem can be reduced to the solvability of a nonlinear elliptic equation. Indeed, if two metrics g_0 and g are related by $g = u^{\frac{4}{n-2}} g_0$ for a smooth positive function u, then the scalar curvature associated with g can be calculated from the scalar curvature associated with g_0 by means of the identity

$$R_g \, u^{\frac{n+2}{n-2}} = R_{g_0} \, u - \frac{4(n-1)}{n-2} \, \Delta_{g_0} u \tag{5}$$

(see [2]). Here, R_{g_0} denotes the scalar curvature associated with g_0, and Δ_{g_0} is the Laplace operator relative to g_0. Hence, if u is a positive solution of the partial differential equation

$$\frac{4(n-1)}{n-2} \, \Delta_{g_0} u - R_{g_0} \, u + r \, u^{\frac{n+2}{n-2}} = 0, \tag{6}$$

then the metric $g = u^{\frac{4}{n-2}} g_0$ has constant scalar curvature r. The solutions of (6) can be characterized as the critical points of the functional

$$E_{g_0}(u) = \frac{\int_M \left(\frac{4(n-1)}{n-2} \, |du|^2_{g_0} + R_{g_0} \, u^2 \right) d\,\mathrm{vol}_{g_0}}{\left(\int_M u^{\frac{2n}{n-2}} \, d\,\mathrm{vol}_{g_0} \right)^{\frac{n-2}{n}}}. \tag{7}$$

The functional $E_{g_0}(u)$ is called the Yamabe energy. It follows from (5) that $E_{g_0}(u) = \mathcal{E}(u^{\frac{4}{n-2}}\, g_0)$, where

$$\mathcal{E}(g) = \frac{\int_M R_g \, d\mathrm{vol}_g}{\big(\int_M d\mathrm{vol}_g\big)^{\frac{n-2}{n}}} \tag{8}$$

denotes the normalized Einstein–Hilbert action. The Yamabe constant of a metric g_0 is defined as the infimum of the Yamabe energy in the conformal class of g_0:

$$Y(M, g_0) = \inf_{0<u\in C^\infty(M)} E_{g_0}(u). \tag{9}$$

The Yamabe constant is closely related to the optimal constant in the Sobolev embedding of $W^{1,2}(M, g_0)$ into $L^{\frac{2n}{n-2}}(M, g_0)$. It is not difficult to show that $Y(M, g_0) \leq Y(S^n)$, where $Y(S^n) = n(n-1)\,\omega_n^{\frac{2}{n}}$ denotes the Yamabe constant of the standard metric on S^n. The key step in the proof of Theorem 2.1 is the following result, which is due to Aubin and Schoen:

Theorem 2.2. *Let M be a compact manifold of dimension $n \geq 3$ without boundary, and let g_0 be a Riemannian metric on M. Suppose that (M, g_0) is not conformally equivalent to the standard sphere S^n. Then $Y(M, g_0) < Y(S^n)$.*

Aubin proved Theorem 2.2 under the additional assumption that $n \geq 6$ and (M, g_0) is not locally conformally flat. The remaining cases were solved by Schoen using the positive mass theorem.

3. The Yamabe flow

R. Hamilton proposed a heat flow approach to the Yamabe problem [15]. Hamilton considered the following evolution equation for the Riemannian metric g:

$$\frac{\partial}{\partial t} g(t) = -(R_{g(t)} - r_{g(t)})\, g(t). \tag{10}$$

Here, $R_{g(t)}$ denotes the scalar curvature associated with the metric $g(t)$. Moreover, the normalization constant $r_{g(t)}$ is defined as the mean value of the scalar curvature on M:

$$r_{g(t)} = \frac{\int_M R_{g(t)} \, d\mathrm{vol}_{g(t)}}{\int_M d\mathrm{vol}_{g(t)}}.$$

This choice of the normalization constant ensures that the volume of M does not change under the evolution. Note that the value of the normalization constant $r_{g(t)}$ may change during the evolution.

The evolution equation (10) can be viewed as a generalization of the Ricci flow in dimension 2 and is often referred to as the Yamabe flow. Like the Ricci flow in dimension 2, the Yamabe flow can be reduced to a parabolic equation for a scalar

function. To see this, we fix a background metric g_0 in the conformal class of the initial metric. Since the Yamabe flow preserves the conformal structure, we may write the time-dependent metric in the form $g(t) = u(t)^{\frac{4}{n-2}}\, g_0$, where $u(t)$ is a positive function on M. Using the relation (5), one can show that the function u satisfies the partial differential equation

$$\frac{\partial}{\partial t} u^{\frac{n+2}{n-2}} = \frac{n+2}{4} \left(\frac{4(n-1)}{n-2}\, \Delta_{g_0} u - R_{g_0}\, u + r_g\, u^{\frac{n+2}{n-2}} \right), \tag{11}$$

which can be viewed as a parabolic analogue of the Yamabe equation (6). It follows from standard parabolic regularity theory that the Yamabe flow has a smooth solution which is defined on a small time interval. Hamilton proved that the Yamabe flow has a global solution for every initial metric [15]. Moreover, he showed that the scalar curvature satisfies the pointwise bound $|R_{g(t)}| \leq C$, where C is a constant that depends only on the initial metric (but not on t).

The discussion of the asymptotic behavior of the flow can be divided into two cases which are distinguished by the sign of the Yamabe constant $Y(M, g_0)$. If $Y(M, g_0) \leq 0$, one can apply the maximum principle to show that the Yamabe flow converges exponentially to a metric of constant scalar curvature.

One of the first results in the case of positive Yamabe constant is due to B. Chow [10]. Chow showed that the flow converges to a metric of constant sectional curvature provided that (M, g_0) is locally conformally flat and the initial metric has positive Ricci curvature. The proof is inspired by Hamilton's work on the Ricci flow in dimension 3 (see [13]), and is based on pinching estimates for the eigenvalues of the Ricci tensor.

R. Ye [43] proved the convergence of the flow for all initial metrics, assuming only that (M, g_0) is locally conformally flat. To this end, he established a gradient bound for the Yamabe flow on locally conformally flat manifolds. The proof of the gradient estimate is based on the method of moving planes and uses the injectivity of the developing map (see [34]).

A different approach was developed by H. Schwetlick and M. Struwe [36]. Among other things, Schwetlick and Struwe proved that the scalar curvature approaches a constant in the sense that

$$\lim_{t \to \infty} \int_M |R_{g(t)} - r_{g(t)}|^p \, d\mathrm{vol}_{g(t)} = 0 \tag{12}$$

for all $p \geq 1$. Using (12), they showed that the Yamabe flow cannot form a singularity unless volume concentration occurs. Moreover, if volume concentration occurs, then the metric can be rescaled in such a way that the rescaled metrics converge to the standard metric on S^n.

Lemma 3.1. *Let $\{t_\nu : \nu \in \mathbb{N}\}$ be a sequence of times such that $t_\nu \to \infty$ as $\nu \to \infty$. Moreover, let $u_\nu = u(t_\nu)$. After passing to a subsequence if necessary, we can find a non-negative integer m and sequences $\{x_{k,\nu} : \nu \in \mathbb{N}\}$, $\{\varepsilon_{k,\nu} : \nu \in \mathbb{N}\}$, $k = 1, \ldots, m$,*

such that

$$u_\nu(x) - \sum_{k=1}^{m} \varphi_{x_{k,\nu}}(x) \left(\frac{4n(n-1)}{r_\infty}\right)^{\frac{n-2}{4}} \left(\frac{\varepsilon_{k,\nu}}{\varepsilon_{k,\nu}^2 + d(x_{k,\nu}, x)^2}\right)^{\frac{n-2}{2}} \to u_\infty(x)$$

in $W^{1,2}(M, g_0)$. *Here,* r_∞ *is defined as the limit of the normalization constant* $r_{g(t)}$ *as* $t \to \infty$. *The function* u_∞ *is a non-negative smooth solution of the partial differential equation*

$$\frac{4(n-1)}{n-2} \Delta_{g_0} u_\infty - R_{g_0}\, u_\infty + r_\infty\, u_\infty^{\frac{n+2}{n-2}} = 0.$$

Finally, $\varphi_{x_{k,\nu}}$ *is a cutoff function such that* $\varphi_{x_{k,\nu}}(x) = 1$ *for* $d(x_{k,\nu}, x) \leq \delta$ *and* $\varphi_{x_{k,\nu}}(x) = 0$ *for* $d(x_{k,\nu}, x) \geq 2\delta$.

This result is due to Schwetlick and Struwe (see [36], Lemma 3.4, and [38]). The proof uses a theorem due to M. Obata, which asserts that every conformal metric on S^n with constant scalar curvature has constant sectional curvature (see [12] or [26]).

Moreover, Schwetlick and Struwe were able to rule out volume concentration provided that $n \leq 5$ and the Yamabe energy of the initial metric is below a certain threshold. The latter condition precludes the formation of a singularity with more than one "bubble". The formation of a singularity with exactly one "bubble" can be ruled out by means of the positive mass theorem.

It was shown in [7] that the condition $n \leq 5$ implies the convergence of the Yamabe flow for all initial metrics (regardless of their Yamabe energy):

Theorem 3.2. *Let* M *be a compact manifold of dimension* $n \geq 3$ *without boundary, and let* g_0 *be a Riemannian metric on* M. *Suppose that* $n \leq 5$ *or* (M, g_0) *is locally conformally flat. Given any initial metric in the conformal class of* g_0, *the Yamabe flow has a global solution which converges to a metric of constant scalar curvature as* $t \to \infty$.

In the locally conformally flat case, this provides an alternative proof of Ye's theorem. The proof of Theorem 3.2 rests on the following key lemma:

Lemma 3.3. *Suppose that* $n \leq 5$ *or* (M, g_0) *is locally conformally flat. Moreover, let* $\{g(t) : t \geq 0\}$ *be a solution of the Yamabe flow on* M. *Then there exist positive real numbers* γ, C, *and* t_0 *such that*

$$r_{g(t)} - r_\infty \leq C \left(\int_M |R_{g(t)} - r_{g(t)}|^{\frac{2n}{n+2}}\, d\mathrm{vol}_{g(t)}\right)^{\frac{n+2}{2n}(1+\gamma)} \tag{13}$$

for $t \geq t_0$.

Suppose for simplicity that the volume is normalized to 1. Then the number $r_{g(t)}$ coincides with the Yamabe energy at time t. Hence, the number r_∞ can be interpreted

as the limit of the Yamabe energy as $t \to \infty$. The difference $r_{g(t)} - r_\infty$ can be expressed as a space-time integral

$$r_{g(t)} - r_\infty = \frac{n-2}{2} \int_t^\infty \int_M (R_{g(\tau)} - r_{g(\tau)})^2 \, d\mathrm{vol}_{g(\tau)} \, d\tau. \tag{14}$$

Hence, we can use (13) and Hölder's inequality to derive a differential inequality for the function $r_{g(t)} - r_\infty$. This implies

$$\int_0^\infty \left(\int_M (R_{g(t)} - r_{g(t)})^2 \, d\mathrm{vol}_{g(t)} \right)^{\frac{1}{2}} dt < \infty. \tag{15}$$

This estimate can be used to rule out volume concentration. More precisely, given any positive real number η, we can find a real number $r > 0$ such that

$$\sup_{x \in M, \, t \geq 0} \int_{B_r(x)} u(t)^{\frac{2n}{n-2}} \, d\mathrm{vol}_{g_0} \leq \eta,$$

where $B_r(x)$ denotes a geodesic ball in the background metric g_0. In light of the results of Schwetlick and Struwe, it follows that

$$\sup_{x \in M, \, t \geq 0} u(x,t) < \infty.$$

Thus, the Yamabe flow converges to a metric of constant scalar curvature as $t \to \infty$. Moreover, if (13) holds with $\gamma = 1$, then the flow converges to the limiting metric at an exponential rate.

To prove Lemma 3.3, we need to find a positive real number ε_0 and a family of auxiliary functions $\bar{u}_{(p,\varepsilon)}$ ($p \in M$, $0 < \varepsilon < \varepsilon_0$) such that

$$E_{g_0}(\bar{u}_{(p,\varepsilon)}) \leq Y(S^n) \tag{16}$$

and $\bar{u}_{(p,\varepsilon)}$ has the "right" asymptotic behavior as $\varepsilon \to 0$. For $n \leq 5$ the existence of such a family of test functions follows from work of R. Schoen [29]. The same approach works if (M, g_0) is locally conformally flat.

Another ingredient in the proof of Lemma 3.3 is an inequality for real-analytic functions due to Lojasiewicz. This inequality was used in the work of L. Simon on the asymptotic behavior of gradient flows [37]. The Lojasiewicz inequality is typically applied to prove the uniqueness of the asymptotic limit of a gradient flow once a-priori estimates have been established.

We cannot, in general, expect (13) to be true for $\gamma = 1$. Indeed, if (13) holds for $\gamma = 1$, then the flow converges to the limiting metric at an exponential rate. This is unlikely to be true in the presence of degenerate solutions.

4. Convergence of the Yamabe flow in dimension greater or equal to 6

We next consider the case $n \geq 6$. In order to prove that the Yamabe flow approaches a metric of constant scalar curvature as $t \to \infty$, we need to construct a family of test functions $\bar{u}_{(p,\varepsilon)}$ that satisfy the inequality (16) among other technical conditions. To this end, it is natural to impose conditions on the Weyl conformal curvature tensor. For example, if we assume that $|W(p)| > 0$ for all $p \in M$, then we can use a result due to Aubin [1] to prove the existence of a family of auxiliary functions with the desired properties.

More generally, suppose that p is a point on M, and let $d = [\frac{n-2}{2}]$ be the largest integer less than or equal to $\frac{n-2}{2}$. If the Weyl tensor does not vanish to an order greater than $d-2$ at p, then we can take advantage of the local geometry to construct a test function with Yamabe energy less than $Y(S^n)$. On the other hand, if the Weyl tensor vanishes to an order greater than $d-2$ at p, we expect that the positive mass theorem can be used to push the energy of the test function below $Y(S^n)$. This motivates the following definition.

Definition 4.1. Let l be a positive integer. We denote by $Z_l(g_0)$ the set of all points $p \in M$ such that

$$\lim_{x \to p} d(x, p)^{2-l} \, |W(x)| = 0,$$

where $W(x)$ denotes the Weyl tensor associated with the metric g_0, and $d(\cdot, \cdot)$ denotes the geodesic distance relative to that metric.

Observe that the set $Z_l(g_0)$ depends only on the conformal class of g_0. It is easy to see that $Z_l(g_0)$ is a compact subset of M. Moreover, we have $M = Z_1(g_0) \supset Z_2(g_0) \supset \cdots$.

Theorem 4.2. *Suppose that $n \geq 6$ and $Z_d(g_0) = \emptyset$ for $d = [\frac{n-2}{2}]$. Then, for every initial metric in the conformal class of g_0, the Yamabe flow has a global solution which approaches a metric of constant scalar curvature as $t \to \infty$.*

To prove Theorem 4.2, we consider an arbitrary point $p \in M$. By a result of J. Lee and T. Parker, we can find a metric g in the conformal class of g_0 such that

$$\det g(x) = 1 + O(|x|^{2d+2}) \tag{17}$$

in geodesic normal coordinates around p (see [21] or [35]). This is called the conformal normal coordinate system. Working in conformal normal coordinates serves two purposes: first, the condition (17) allows us to simplify some of the calculations. Second, it can be shown that the metric agrees with the flat metric to the maximal order permitted by the Weyl tensor. More precisely, if $p \in Z_d(g_0)$, then $g_{ik}(x) = \delta_{ik} + O(|x|^{d+1})$ in conformal normal coordinates. (Conversely, if $g_{ik}(x) = \delta_{ik} + O(|x|^{d+1})$ for any metric conformally equivalent to g_0, then clearly p belongs to the set $Z_d(g_0)$.)

It is convenient to write the Riemannian metric in the form $g(x) = \exp(h(x))$, where $h(x)$ is a symmetric 2-tensor satisfying

$$\operatorname{tr} h(x) = O(|x|^{2d+2}). \tag{18}$$

We denote by $H_{ik}(x) = \sum_{2 \le |\alpha| \le d} h_{ik,\alpha}\, x^\alpha$ the Taylor polynomial of order d associated with the function $h_{ik}(x)$. Since $p \notin Z_d(g_0)$, at least one of the polynomials $H_{ik}(x)$ $(1 \le i, k \le n)$ is not identically zero.

Given a positive real number ε, we define a function $u_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by

$$u_\varepsilon(x) = \left(\frac{\varepsilon}{\varepsilon^2 + |x|^2}\right)^{\frac{n-2}{2}},$$

so that

$$\Delta u_\varepsilon + n(n-2)\, u_\varepsilon^{\frac{n+2}{n-2}} = 0. \tag{19}$$

Our aim is to construct a test function $\bar{u}_{(p,\varepsilon)}$ which is close to u_ε and has Yamabe energy less than $Y(S^n)$. To this end, we exploit the saddle point structure of the Einstein–Hilbert action near the standard metric on S^n (see [6], Section 4G). We first choose a vector field $V^\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\sum_{k=1}^n \partial_k \Big[u_\varepsilon^{\frac{2n}{n-2}} \left(H_{ik} - \partial_i V_k^\varepsilon - \partial_k V_i^\varepsilon + \frac{2}{n} \operatorname{div} V^\varepsilon\, \delta_{ik} \right) \Big] = 0 \tag{20}$$

for $i = 1, \dots, n$. This is a linear elliptic system. In order to construct a solution to (20), it suffices to minimize the functional

$$\int_{\mathbb{R}^n} u_\varepsilon^{\frac{2n}{n-2}} \sum_{i,k=1}^n \left(H_{ik} - \partial_i V_k - \partial_k V_i + \frac{2}{n} \operatorname{div} V\, \delta_{ik} \right)^2$$

over all vector fields $V : \mathbb{R}^n \to \mathbb{R}^n$.

In the next step, we define a function $v_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by

$$v_\varepsilon = \sum_{k=1}^n \partial_k u_\varepsilon\, V_k^\varepsilon + \frac{n-2}{2n}\, u_\varepsilon \operatorname{div} V^\varepsilon.$$

It follows from (20) that the function v_ε solves the linearized equation

$$\Delta v_\varepsilon + n(n+2)\, u_\varepsilon^{\frac{4}{n-2}}\, v_\varepsilon = \sum_{i,k=1}^n \frac{n-2}{4(n-1)}\, u_\varepsilon\, \partial_i \partial_k H_{ik}. \tag{21}$$

The fact that v_ε is a solution of the linearized equation (21) suggests that $u_\varepsilon + v_\varepsilon$ is a

good candidate for a test function. Indeed, one can show that

$$\begin{aligned}&\int_{B_\delta(0)} \Big(\frac{4(n-1)}{n-2}\,|d(u_\varepsilon+v_\varepsilon)|_g^2 + R_g\,(u_\varepsilon+v_\varepsilon)^2\Big) \\ &\quad\leq \bigg(\int_{B_\delta(0)} (u_\varepsilon+v_\varepsilon)^{\frac{2n}{n-2}}\bigg)^{\frac{n-2}{n}} Y(S^n) \\ &\qquad - \theta \sum_{2\leq|\alpha|\leq d} \sum_{i,k=1}^n |h_{ik,\alpha}|^2\,\varepsilon^{2|\alpha|} + C(\delta)\,\varepsilon^{n-2}\end{aligned} \tag{22}$$

if n is odd. A similar estimate holds if n is even: in this case, we have

$$\begin{aligned}&\int_{B_\delta(0)} \Big(\frac{4(n-1)}{n-2}\,|d(u_\varepsilon+v_\varepsilon)|_g^2 + R_g\,(u_\varepsilon+v_\varepsilon)^2\Big) \\ &\quad\leq \bigg(\int_{B_\delta(0)} (u_\varepsilon+v_\varepsilon)^{\frac{2n}{n-2}}\bigg)^{\frac{n-2}{n}} Y(S^n) \\ &\qquad - \theta \sum_{2\leq|\alpha|\leq d-1} \sum_{i,k=1}^n |h_{ik,\alpha}|^2\,\varepsilon^{2|\alpha|} \\ &\qquad - \theta \sum_{|\alpha|=d} \sum_{i,k=1}^n |h_{ik,\alpha}|^2\,\varepsilon^{n-2}\,\log\frac{1}{\varepsilon} + C(\delta)\,\varepsilon^{n-2}.\end{aligned} \tag{23}$$

In both inequalities, θ is a positive constant that depends only on n. It remains to extend the function $u_\varepsilon + v_\varepsilon$ to all of M. In view of (22) and (23), this can be done in such a way that the Yamabe energy of the resulting function is less than $Y(S^n)$ if ε is sufficiently small. The details will appear in a forthcoming paper.

We expect that the conclusion of Theorem 4.2 is still valid without the condition $Z_d(g_0) = \emptyset$. A proof of this conjecture will likely involve the positive mass theorem. The positive mass theorem was first proved in dimension 3 by R. Schoen and S.T. Yau using minimal surface techniques [33]. This argument can be extended up to dimension 7, cf. [31]. E. Witten gave an alternative proof of the positive mass theorem based on spinor methods [41]. (This approach works only for spin manifolds.) R. Bartnik extended Witten's arguments to prove the positive mass theorem for spin manifolds of any dimension [4]. J. Lohkamp recently announced a proof of the positive mass theorem in arbitrary dimension, which does not require the manifold to be spin [24].

5. Compactness of the set of constant scalar curvature metrics in a given conformal class

In this section, I will review several recent results concerning the set of solutions to (6). In particular, I will discuss under what conditions the set of solutions is compact. These

results are remarkably similar to those stated in the preceding sections. As above, we assume that M is a compact manifold of dimension $n \geq 3$ without boundary.

Theorem 5.1. *Suppose that $n \leq 7$ or (M, g_0) is locally conformally flat. Moreover, we assume that (M, g_0) is not conformally equivalent to the standard sphere S^n. Let r be a fixed real number, and let u be a positive solution of* (6). *Then there exists a constant C, depending only on g_0 and r, such that $\|u\|_{C^2(M,g_0)} \leq C$ and $\inf_M u \geq 1/C$.*

This compactness result was proved by R. Schoen in [30] (see also [31] and [32]). Y. Y. Li and M. Zhu gave an alternative proof in dimension 3 [23]. The extension to dimensions 4 and 5 is due to O. Druet [11]. The cases $n = 6$ and $n = 7$ were solved independently by Y.Y. Li and L. Zhang [22] and F. Marques [25].

M. Khuri and R. Schoen [20] recently established a compactness result in arbitrary dimension. Among other things, Khuri and Schoen proved that the Weyl tensor vanishes to an order greater than $[\frac{n-6}{2}]$ at each blow-up point. In particular, if $Z_d(g_0) = \emptyset$ for $d = [\frac{n-2}{2}]$, then blow-up cannot occur.

Theorem 5.2. *Suppose that $n \geq 8$ and $Z_d(g_0) = \emptyset$ for $d = [\frac{n-2}{2}]$. Moreover, let r be a fixed real number. Then there exists a constant C, depending only on g_0 and r, such that $\|u\|_{C^2(M,g_0)} \leq C$ and $\inf_M u \geq 1/C$ for every positive solution of* (6).

A special case of Theorem 5.2 was proved in a recent paper by Y. Y. Li and L. Zhang [22]. Li and Zhang assumed that $n \geq 8$ and $|W(p)| + |\nabla W(p)| > 0$ for all $p \in M$. (This condition is equivalent to $Z_3(g_0) = \emptyset$.) The proof of the main result in [22] uses special properties of conformal normal coordinates established by E. Hebey and M. Vaugon [17].

Moreover, Khuri and Schoen showed that the condition $Z_d(g_0) = \emptyset$ can be removed by means of the positive mass theorem. Since the positive mass theorem holds for spin manifolds of any dimension, this yields the following result:

Theorem 5.3. *Suppose that M is a spin manifold and (M, g_0) is not conformally equivalent to the standard sphere S^n. Moreover, suppose that u is a positive solution of* (6) *for some fixed real number r. Then $\|u\|_{C^2(M,g_0)} \leq C$ and $\inf_M u \geq 1/C$, where C depends only on g_0 and r.*

It is shown in [20] that the compactness result remains true if the equation (6) is replaced by a family of subcritical equations. This is useful in some applications. For example, this can be used to obtain results concerning the number of constant scalar curvature metrics in a given conformal class.

It is interesting to compare the results above to the following result of D. Pollack (see [28], Theorem 0.1):

Theorem 5.4. *Suppose that $Y(M, g_0) > 0$. Given any positive integer N, there exists a Riemannian metric g with the following properties:*

(i) $\|g - g_0\|_{C^0(M,g_0)} \leq 1/N$.

(ii) *The equation* $\frac{4(n-1)}{n-2}\,\Delta_g u - R_g\,u + u^{\frac{n+2}{n-2}} = 0$ *has at least N positive solutions.*

It follows from the results mentioned above that Theorem 5.4 cannot hold if the C^0-norm is replaced by the C^l-norm for a sufficiently large integer l (which may depend on the dimension n). The reason is that the a-priori estimates for solutions of (6) are stable under perturbations of the background metric that are small in the C^l-topology.

References

[1] Aubin, T., Équations différentielles non linéaires et problème de Yamabe concernant la courbure scalaire. *J. Math. Pures Appl.* **55** (1976), 269–296.

[2] Aubin, T., *Some nonlinear problems in Riemannian geometry.* Springer Monogr. Math., Springer-Verlag, Berlin 1998.

[3] Bahri, A., Proof of the Yamabe conjecture, without the positive mass theorem, for locally conformally flat manifolds. In *Einstein metrics and Yang-Mills connections* (ed. by Toshiki Mabuchi and Shigeru Mukai), Lecture Notes Pure Appl. Math. 145, Marcel Dekker, New York 1993, 1–26.

[4] Bartnik, R., The mass of an asymptotically flat manifold. *Comm. Pure Appl. Math.* **39** (1986), 661–693.

[5] Berger, M., On Riemannian structures of prescribed Gaussian curvature for compact 2-surfaces. *J. Differential Geometry* **5** (1971), 325–332.

[6] Besse, A. L., *Einstein manifolds.* Ergeb. Math. Grenzgeb. (3) 10, Springer-Verlag, Berlin 1987.

[7] Brendle, S., Convergence of the Yamabe flow for arbitrary initial energy. *J. Differential Geometry* **69** (2005), 217–278.

[8] Chen, X., Calabi flow in Riemann surfaces revisited: a new point of view. *Internat. Math. Res. Notices* **6** (2001), 275–297.

[9] Chow, B., The Ricci flow on the 2-sphere. *J. Differential Geometry* **33** (1991), 325–334.

[10] Chow, B., The Yamabe flow on locally conformally flat manifolds with positive Ricci curvature. *Comm. Pure. Appl. Math.* **45** (1992), 1003–1014.

[11] Druet, O., Compactness for Yamabe metrics in low dimensions. *Internat. Math. Res. Notices* **23** (2004), 1143–1191.

[12] Gidas, B., Ni, W., Nirenberg, L., Symmetry and related properties via the maximum principle. *Comm. Math. Phys.* **68** (1979), 209–243.

[13] Hamilton, R. S., Three-manifolds with positive Ricci curvature. *J. Differential Geometry* **17** (1982), 255–306.

[14] Hamilton, R. S., The Ricci flow on surfaces. In *Mathematics and general relativity* (ed. by James Isenberg), Contemp. Math. 71, Amer. Math. Soc., Providence, RI, 1988, 237–262.

[15] Hamilton, R. S., Lectures on geometric flows. Unpublished, 1989.

[16] Hamilton, R. S., An isoperimetric estimate for the Ricci flow on the two-sphere. In *Modern methods in complex analysis* (ed. by Thomas Bloom, David Catlin, John P. D'Angelo and Yum-Tong Siu), Ann. of Math. Stud. 137, Princeton University Press, Princeton, NJ, 1995, 191–200.

[17] Hebey, E., Vaugon, M., Le problème de Yamabe équivariant. *Bull. Sci. Math.* **117** (1993), 241–286.

[18] Kazdan, J., Warner, F., Curvature functions of compact 2-manifolds. *Ann. of. Math.* **99** (1974), 14–47.

[19] Kazdan, J., Warner, F., Existence and conformal deformation of metrics with prescribed Gaussian and scalar curvatures. *Ann. of Math.* **101** (1975), 317–331.

[20] Khuri, M., Schoen, R. M., private communication.

[21] Lee, J. M., Parker, M., The Yamabe problem. *Bull. Amer. Math. Soc.* **17** (1987), 37–91.

[22] Li, Y. Y., Zhang, L., Compactness of solutions to the Yamabe problem II. *Calc. Var. Partial Differential Equations* **24** (2005), 185–237.

[23] Li, Y. Y., Zhu, M., Yamabe type equations on three dimensional Riemannian manifolds. *Commun. Contemp. Math.* **1** (1999), 1–50.

[24] Lohkamp, J., private communication.

[25] Marques, F. C., A-priori estimates for the Yamabe problem in the non-locally conformally flat case. math.DG/0408063.

[26] Obata, M., The conjectures on conformal transformations of Riemannian manifolds. *J. Differential Geometry* **6** (1971), 247–258.

[27] Osgood, B., Phillips, R., Sarnak, P., Extremals of determinants of Laplacians. *J. Funct. Anal.* **80** (1988), 148–211.

[28] Pollack, D., Nonuniqueness and high energy solutions for a conformally invariant scalar equation. *Comm. Anal. Geom.* **1** (1993), 347–414.

[29] Schoen, R. M., Conformal deformation of a Riemannian metric to constant scalar curvature. *J. Differential Geometry* **20** (1984), 479–495.

[30] Schoen, R. M., Courses taught at Stanford University, 1988.

[31] Schoen, R. M., Variational theory for the total scalar curvature functional for Riemannian metrics and related topics. In *Topics in the calculus of variations* (ed. by Mariano Giaquinta), Lecture Notes in Math. 1365, Springer-Verlag, Berlin 1989, 120–154.

[32] Schoen, R. M., On the number of constant scalar curvature metrics in a conformal class. In *Differential geometry* (ed. by H. Blaine Lawson, Jr., and Keti Tenenblat), Pitman Monogr. Surveys Pure Appl. Math. 52, Longman Scientific & Technical, Harlow 1991, 311–320.

[33] Schoen, R. M., Yau, S. T., On the proof of the positive mass conjecture in general relativity. *Comm. Math. Phys.* **65** (1979), 45–76.

[34] Schoen, R. M., Yau, S. T., Conformally flat manifolds, Kleinian groups and scalar curvature. *Invent. Math.* **92** (1988), 47–71.

[35] Schoen, R. M., Yau, S. T., *Lectures on differential geometry*. Conf. Proc. Lecture Notes Geom. Topology 1, International Press, Cambridge, MA, 1994.

[36] Schwetlick, H., Struwe, M., Convergence of the Yamabe flow for large energies. *J. Reine Angew. Math.* **562** (2003), 59–100.

[37] Simon, L., Asymptotics for a class of non-linear evolution equations with applications to geometric problems. *Ann. of. Math.* **118** (1983), 525–571.

[38] Struwe, M., A global compactness result for elliptic boundary value problems involving limiting nonlinearities. *Math. Z.* **187** (1984), 511–517.

[39] Struwe, M., Curvature flows on surfaces. *Ann. Scuola Norm. Sup. Pisa* (5) **1** (2002), 247–274.

[40] Trudinger, N., Remarks concerning the conformal deformation of Riemannian structures on compact manifolds. *Ann. Scuola Norm. Sup. Pisa* **22** (1968), 265–274.

[41] Witten, E., A new proof of the positive energy theorem. *Comm. Math. Phys.* **80** (1981), 381–402.

[42] Yamabe, H., On a deformation of Riemannian structures on compact manifolds. *Osaka Math J.* **12** (1960), 21–37.

[43] Ye, R., Global existence and convergence of the Yamabe flow. *J. Differential Geometry* **39** (1994), 35–50.

Stanford University, Department of Mathematics, 450 Serra Mall, Bldg. 380, Stanford, CA 94305, U.S.A.

E-mail: brendle@math.stanford.edu

The topology and geometry of contact structures in dimension three

Ko Honda *

Abstract. The goal of this article is to survey recent developments in the theory of contact structures in dimension three.

Mathematics Subject Classification (2000). Primary 57M50; Secondary 53C15.

Keywords. Tight, contact structure, bypass, open book decomposition, mapping class group, Dehn twists, Reeb vector field, contact homology.

In this article we survey recent developments in three-dimensional contact geometry. Three-dimensional contact geometry lies at the interface between 3- and 4-manifold geometries, and has been an essential part of the flurry in low-dimensional geometry and topology over the last 20 years. In dimension 3, it relates to foliation theory and knot theory; in dimension 4, there are rich interactions with symplectic geometry. In both dimensions, there are relations with gauge theories such as Seiberg–Witten theory and Heegaard–Floer homology, as well as to dynamics.

1. Tight vs. overtwisted

A contact structure ξ on a 3-manifold M is a (maximally) nonintegrable 2-plane field distribution. In this paper we assume that M is oriented and ξ is the kernel of a global 1-form α which satisfies $\alpha \wedge d\alpha > 0$. (Such a contact structure is often called *coorientable*.) Although ξ is locally $\mathrm{Ker}(dz - ydx)$ by a classical theorem of Pfaff–Darboux and hence has *no local geometry*, the global study of contact structures is rather complicated, in a way that echoes the intricacies of symplectic geometry.

One of the fundamental questions is to determine π_0 of the space $\mathrm{Cont}(M)$ of contact 2-plane fields on M – this is often called the "classification" of contact structures on M. The work of Bennequin [2] (later clarified by Eliashberg [12]) indicated that contact structures, in dimension three, come in two flavors: *tight* and *overtwisted*. We define an *overtwisted disk* to be an embedded disk $D \subset M$ such that $\xi_x = T_x D$ for all $x \in \partial D$. An overtwisted contact structure is one which admits an *overtwisted disk*,

*Supported by an Alfred P. Sloan Fellowship and an NSF CAREER Award.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

whereas a tight contact structure is one which does not. What Bennequin showed is that the local model $(\mathbb{R}^3, \mathrm{Ker}(dz - ydx))$ is tight – hence locally every contact structure is tight, although globally it may not be. (Showing that a contact structure is tight is highly nontrivial, because one must show that *no* overtwisted disk exists, no matter how complicated the embedding!) Let $\mathrm{Cont}^{\mathrm{OT}}(M)$ be the space of overtwisted contact 2-plane fields on M and $\mathrm{Cont}^{\mathrm{Tight}}(M)$ be the space of tight contact 2-plane fields on M. Eliashberg showed in [13] that $\pi_0(\mathrm{Cont}^{\mathrm{OT}}(M))$ is the same as the homotopy classes of 2-plane fields on M.

The space of tight contact structures, on the other hand, is more intimately related to the topology of M. Eliashberg [12] gave the first classification result for tight contact structures, namely that $\mathrm{Cont}^{\mathrm{Tight}}(S^3)$ is connected. The analysis of tight contact structures on various 3-manifolds has become more manageable in recent years, with the introduction of *convex surfaces* by Giroux [32] and *bypasses* by the author [43]. The world of tight contact structures, as we understand it now, is a veritable zoo!

Two important subcategories of tight contact structures are the *weakly symplectically fillable* ones and the *Stein fillable* ones. In the former case, (M, ξ) bounds a symplectic 4-manifold (X, ω) and $\omega|_\xi > 0$. In the latter, (M, ξ) bounds a Stein domain (X, ω, J) and $\omega = d\alpha$ on M for a contact 1-form α that defines ξ. Fillable contact structures (of either type) are tight by a theorem of Gromov [36] and Eliashberg [14]; this was proved using Gromov's theory of J-holomorphic curves. Prototypical examples of weakly symplectically fillable contact structures are the perturbations of taut codimension 1 foliations, as explained in Eliashberg–Thurston [19]. Etnyre and the author [23] showed that there exist tight contact structures which are not weakly symplectically fillable. Other examples were later obtained by Lisca–Stipsicz [51], [52]. Eliashberg [15] showed that there are weakly symplectically fillable contact structures on the 3-torus T^3 which are not Stein fillable. Further examples were given on torus bundles by Ding–Geiges [11].

$$\text{Tight} \supsetneq \text{Weakly symplectically fillable} \supsetneq \text{Stein fillable}$$

It is known that not every 3-manifold admits a tight contact structure. Etnyre and the author [22] showed that the Poincaré homology sphere with orientation opposite to the one induced on the link of an algebraic singularity has no tight contact structure. Lisca and Stipsicz [54] have since shown that the Poincaré homology sphere can be incorporated into a larger class of small Seifert fibered spaces which do not admit tight contact structures. Since all these examples of 3-manifolds without tight contact structures are Seifert fibered, it is natural to ask whether tight contact structures exist on all hyperbolic 3-manifolds. It turns out that *universally tight* contact structures, i.e., contact structures ξ on M that pull back to tight contact structures on the universal cover of M, do not always exist on hyperbolic 3-manifolds [47]. Compare this to foliation theory where Roberts–Shareshian–Stein [66] have shown that there are

infinitely many hyperbolic 3-manifolds which do not admit Reebless codimension 1 foliations. There is related work of Calegari–Dunfield [4] and Fenley [25], as well as a different approach using Seiberg–Witten Floer homology, due to Kronheimer–Mrowka–Ozsváth–Szabó [50]. However, it is still conceivable that every hyperbolic 3-manifold has a tight contact structure. (The Weeks manifold – the closed hyperbolic 3-manifold with the smallest known volume – *does* have Stein fillable contact structures with either orientation. This can be easily seen by appealing to surgery on the Borromean rings as in Gompf [34].)

Next we turn our attention to the question of classification. For simplicity, assume that M is irreducible. Colin [6] and Kazez, Matić and the author [44], independently, have shown that $\pi_0(\mathrm{Cont}^{\mathrm{Tight}}(M))$ is infinite for a *toroidal* 3-manifold M, namely one which admits an embedded torus for which $\pi_1(T) \hookrightarrow \pi_1(M)$. On the other hand, if M is not toroidal, then $\pi_0(\mathrm{Cont}^{\mathrm{Tight}}(M))$ is finite by a theorem of Colin, Giroux and the author [7], [8]. The latter generalizes an earlier theorem, due to Kronheimer–Mrowka [48], which states that there are finitely many homotopy classes of 2-plane fields which carry symplectically fillable contact structures.

2. Open book decompositions

A fundamental advance in contact geometry is the work of Giroux [33] (building on earlier work of Thurston–Winkelnkemper [69], Bennequin [2], Eliashberg–Gromov [18], and Torisu [70]), which relates contact structures and open book decompositions. We briefly summarize this work, and then describe the developments that have taken place since Giroux's ICM 2002 article [33].

Let (S, h) be a pair consisting of a compact oriented surface S with nonempty boundary and a diffeomorphism $h\colon S \to S$ which restricts to the identity on ∂S, and let K be a link in a closed oriented 3-manifold M. An *open book decomposition* for M with *binding* K is a homeomorphism between $((S \times [0,1])/_{\sim_h}, (\partial S \times [0,1])/_{\sim_h})$ and (M, K). The equivalence relation $\sim_h$ is generated by $(x,1) \sim_h (h(x),0)$ for $x \in S$ and $(y,t) \sim_h (y,t')$ for $y \in \partial S$. We will often identify M with $(S \times [0,1])/_{\sim_h}$; with this identification $S_t = S \times \{t\}$, $t \in [0,1]$, is called a *page* of the open book decomposition and h is called the *monodromy map*. Two open book decompositions are *equivalent* if there is an ambient isotopy taking binding to binding and pages to pages. We will denote an open book decomposition by (S, h), although, strictly speaking, an open book decomposition is determined by the triple (S, h, K). There is a slight difference – if we do not specify $K \subset M$, we are referring to isomorphism classes of open books instead of isotopy classes.

Every closed 3-manifold has an open book decomposition, but it is not unique. One way of obtaining a different open book decomposition of the same manifold is to perform a positive stabilization. (S', h') is a *positive stabilization* of (S, h) if S' is the union of the surface S and a band B attached along the boundary of S (i.e., S' is obtained from S by attaching a 1-handle along ∂S), and h' is defined as follows.

Let γ be a simple closed curve in S' "dual" to the cocore of B (i.e., γ intersects the cocore of B at exactly one point) and let $\mathrm{id}_B \cup h$ be the extension of h by the identity map to $B \cup S$. Also let R_γ be a *positive* or *right-handed* Dehn twist about γ. Then for a *positive* stabilization h' is given by $R_\gamma \circ (\mathrm{id}_B \cup h)$. It is well-known that, if (S', h') is a positive stabilization of an open book decomposition (S, h) of (M, K), then (S', h') is an open book decomposition of (M, K') where K' is obtained by a Murasugi sum of K (also called the *plumbing* of K) with a positive Hopf link.

A contact structure ξ is said to be *supported* by the open book decomposition (S, h, K) if there is a contact 1-form α satisfying the following:

1. $d\alpha$ restricts to a symplectic form on each fiber S_t;
2. K is transverse to ξ, and the orientation on K given by α is the same as the boundary orientation induced from S coming from the symplectic structure.

Thurston and Winkelnkemper [69] showed that any open book decomposition (S, h, K) of M supports a contact structure ξ. Moreover, the contact planes can be made arbitrarily close to the tangent planes of the pages (away from the binding).

The following result provides a converse (and more) due to Giroux [33].

Theorem 2.1 (Giroux). *Every contact structure (M, ξ) on a closed 3-manifold M is supported by some open book decomposition (S, h, K). Moreover, two open book decompositions (S, h, K) and (S', h', K') which support the same contact structure (M, ξ) become equivalent after applying a sequence of positive stabilizations to each.*

2.1. Concave symplectic fillings. Consider a closed 2-form ω_0 on the contact 3-manifold (M, ξ) for which $\omega_0|_\xi > 0$. (Such a 2-form ω_0 is often called a *dominating* 2-form for ξ.) A *concave symplectic filling* for (M, ξ, ω_0) is a symplectic 4-manifold (X, ω) for which $\partial X = -M$ and $i^*\omega = \omega_0$, where $i\colon M \to X$ is the inclusion.

The use of open book decompositions enabled Eliashberg [16] and Etnyre [20] to construct concave symplectic fillings for any contact 3-manifold (M, ξ) together with a dominating 2-form ω_0. This turned out to be the only missing ingredient in Kronheimer–Mrowka's proof of Property P for knots [49].

Theorem 2.2 (Kronheimer–Mrowka). *If $K \subset S^3$ is a nontrivial knot and $S^3_1(K)$ is the three-manifold obtained by $+1$-surgery along K, then $\pi_1(S^3_1(K)) \neq 0$.*

In the 1980's Gabai [27] proved that $M = S^3_0(K)$ admits a taut foliation $\mathcal{F}$ ($\neq$ the foliation of $S^1 \times S^2$ by $\{pt\} \times S^2$) if K is not the unknot. By Eliashberg–Thurston [19], $\mathcal{F}$ can be perturbed into a pair ξ_+, ξ_- of positive and negative contact structures, and $X = M \times [0, 1]$ admits a symplectic structure ω for which $\omega|_{\xi_+} > 0$ at $M \times \{1\}$ and $\omega|_{\xi_-} < 0$ at $M \times \{0\}$. This used to be called a *symplectic semi-filling* of (M, ξ_+) since ∂X had more than one component. (We no longer have the need to use the "semi" terminology, thanks to Eliashberg and Etnyre.) The work of Eliashberg and

Etnyre enabled one to fill both of the components of ∂X so that (M, ξ_+) was now embedded in a closed symplectic manifold X'. Kronheimer and Mrowka were then able to appeal to: (i) the work by Taubes [67] on the nontriviality of Seiberg–Witten invariants of X'; (ii) the work by Feehan and Leness (see [24], for example) relating the Seiberg–Witten and Donaldson invariants; (iii) a stretching argument in instanton Floer homology; and (iv) Floer's exact triangle [26] for instanton Floer homology.

Another application of the existence of concave symplectic fillings is progress by Etnyre [21] on the following problem: Given a contact manifold (M, ξ), what is the minimum genus amongst all the pages of open books corresponding to ξ? Etnyre has shown that many interesting classes of tight contact structures (among them perturbations of taut foliations) do not admit planar open book decompositions.

2.2. Heegaard–Floer homology. Another important application of the open book framework is the definition of the *contact class* $c(\xi)$ in the Heegaard Floer homology of Ozsváth–Szabó [59], [60]. Using open book decompositions, Ozsváth and Szabó [61] defined an invariant of the contact structure (M, ξ), which is an element $c(\xi) \in \widehat{HF}(-M)$. Among the many properties enjoyed by $c(\xi)$, we have the following:

1. If ξ is overtwisted, then $c(\xi) = 0$.

2. If (M', ξ') is obtained from (M, ξ) by Legendrian (-1) surgery, and $c(\xi) \neq 0$, then $c(\xi') \neq 0$.

3. If ξ is weakly symplectically fillable, then $c(\xi) \neq 0$ (provided "twisted" coefficients are used).

Lisca and Stipsicz [53] showed that the contact class was surprisingly good at detecting tight contact structures – Heegaard–Floer homology could now be used to prove the tightness of many contact structures which were hitherto only conjectured to be tight. This area is currently an active area of research, with contributions from Ghiggini [30], [31], Plamenevskaya [62], [63], etc.

3. Right-veering

We will now seek to explain the roles of tightness, weak symplectic fillability, and Stein fillability in the open book context. Except for Theorem 3.1, this is joint work with W. Kazez and G. Matić and further details can be found in [45], [46].

Let S be a compact oriented surface with nonempty boundary and denote by $\mathrm{Aut}(S, \partial S)$ the group of (isotopy classes of) diffeomorphisms of S which restrict to the identity on ∂S. We have the monoid $\mathrm{Dehn}^+(S, \partial S) \subset \mathrm{Aut}(S, \partial S)$ of products of positive Dehn twists. The following is due to Giroux [33], inspired by the work of Loi–Piergallini [55]:

Theorem 3.1 (Giroux). *A contact structure ξ on M is Stein fillable if and only if ξ is supported by some open book (S, h, K) with $h \in \mathrm{Dehn}^+(S, \partial S)$.*

We remark that the theorem does not say that *every* open book (S, h) for (M, ξ) Stein fillable satisfies $h \in \mathrm{Dehn}^+(S, \partial S)$.

There is another monoid, namely the monoid $\mathrm{Veer}(S, \partial S)$ of *right-veering* diffeomorphisms, which is intimately connected with the tight contact structures. Given two properly embedded oriented arcs α and β with the same initial point $x \in \partial S$, we say α is *to the left* of β if the following holds: Isotop α and β, while fixing their endpoints, so that they intersect transversely (this include the endpoints) and with the fewest possible number of intersections. We then say α is to the left of β if either $\alpha = \beta$ or the tangent vectors $(\dot\beta(0), \dot\alpha(0))$ define the orientation on S at x. Then h is *right-veering* if for every choice of basepoint $x \in \partial S$ and every choice of α based at x, $h(\alpha)$ is to the right of α. One easily sees that $\mathrm{Dehn}^+(S, \partial S) \subset \mathrm{Veer}(S, \partial S)$.

Theorem 3.2 (Honda–Kazez–Matić [45]). *A contact structure (M, ξ) is tight if and only if all of its open book decompositions (S, h) are such that $h \in \mathrm{Veer}(S, \partial S)$.*

This theorem is an improvement over the "sobering arc" criterion for overtwistedness, given by Goodman [35].

Now recall Thurston's classification of surface diffeomorphisms [68], which improved upon earlier work of Nielsen [56], [57], [58]. A diffeomorphism $h\colon S \to S$ satisfies one of the following:

1. h is *reducible*, i.e., there exists an essential multicurve γ such that $h(\gamma)$ is isotopic to γ.
2. h is homotopic to a *periodic homeomorphism* ψ, i.e., there is an integer $n > 0$ such that $\psi^n = \mathrm{id}$.
3. h is homotopic to a *pseudo-Anosov homeomorphism* ψ.

We will now define the *fractional Dehn twist coefficients*, extensively studied by Gabai and Oertel (see for example [28]) in the context of essential laminations. Suppose for simplicity that ∂S is connected and $h \in \mathrm{Aut}(S, \partial S)$ is homotopic to a pseudo-Anosov representative ψ. (The periodic case is analogous.) Let $H\colon S \times [0, 1] \to S$ be an isotopy from h to ψ, i.e., $H(x, 0) = h(x)$ and $H(x, 1) = \psi(x)$. On the boundary of S, ψ has $2n$ fixed points, n attracting and n repelling. Let us label the attracting fixed points $x_1, \ldots, x_n$ in order around ∂S. Now define $\beta\colon \partial S \times [0, 1] \to \partial S \times [0, 1]$ by sending $(x, t) \mapsto (H(x, t), t)$. It follows that the arc $\beta(x_i \times [0, 1])$ connects $(x_i, 0)$ and $(x_{i+k}, 1)$, for some k. We call β a *fractional Dehn twist* by an amount $c \in \mathbb{Q}$, where $c \equiv k/n$ modulo 1 is the number of times $\beta(x_i \times [0, 1])$ circles around $\partial S \times [0, 1]$ (here circling in the direction of ∂C is considered positive). Form the union of $\partial S \times [0, 1]$ and S by gluing $\partial S \times \{1\}$ and ∂S. By identifying this union with S, we construct the homeomorphism $\beta \cup \psi$ on S which is isotopic to h, relative

to ∂S. (We will assume that $h = \beta \cup \psi$, although ψ is usually just a homeomorphism, not a diffeomorphism.)

Theorem 3.3 (Honda–Kazez–Matić [45]). *If h is periodic, then h is right-veering if and only if $c \geq 0$. If h is pseudo-Anosov, then h is right-veering if and only if $c > 0$.*

Theorem 3.2 is not completely satisfactory – ideally one should just need to look at one (S, h) (instead of its equivalence class under stabilizations) to determine whether (S, h) is tight, fillable, etc. To this end, let us consider the case of the once-punctured torus S. Suppose h is pseudo-Anosov. Then the following hold:

1. If $c \leq 0$, then h is overtwisted.
2. If $c > 0$, then h is tight.
3. If $c \geq 1$, then h is weakly symplectically fillable and universally tight.
4. For any $c > 0$ there exist $h \in \mathrm{Veer}(S, \partial S) - \mathrm{Dehn}^+(S, \partial S)$ whose fractional Dehn twist coefficient is equal to c.

For the once-punctured torus, $c > 0$ is equivalent to $c \geq \frac{1}{2}$, since the pseudo-Anosov representative ψ will have $n = 2$ attracting fixed points. (2) is proved using Heegaard–Floer homology. (3) is proved by showing that the taut foliations constructed by Roberts in [64], [65] can be perturbed to the contact structure adapted to the open book. (4) is proved by looking at a function on the Farey tessellation called the *Rademacher function* (see for example [29]). It is very plausible that many, if not all, of the $h \in \mathrm{Veer}(S, \partial S) - \mathrm{Dehn}^+(S, \partial S)$ never become products of positive Dehn twists after (repeated) stabilization, and that such (S, h) are indeed not Stein fillable.

Once we restrict our attention to right-veering (S, h), calculations in *contact homology* and *Heegaard–Floer homology* both become more manageable. In the rest of the paper, we focus on contact homology, leaving the Heegaard–Floer aspects for another occasion.

4. Contact homology

Given a contact form α for (M, ξ), there is a corresponding *Reeb vector field* R defined as follows: $i_R d\alpha = 0$, $i_R \alpha = 1$. One of the motivating questions in the study of the dynamics of Reeb vector fields is the following Weinstein conjecture (in dimension three):

Conjecture 4.1 (Weinstein conjecture). Let (M, ξ) be a contact 3-manifold. Then for any contact form α with $\mathrm{Ker}(\alpha) = \xi$, the corresponding Reeb vector field R admits a closed periodic orbit.

The fundamental step was taken when Hofer [37] studied J-holomorphic disks in the symplectization $(\mathbb{R} \times M, d(e^t\alpha))$, and proved that there is always bubbling (and hence a closed periodic orbit) when (M, ξ) is overtwisted or $\pi_2(M) \neq 0$. This showed that the Weinstein conjecture holds for overtwisted contact structures and 3-manifolds with $\pi_2(M) \neq 0$. (Hofer also showed that the Weinstein conjecture holds for S^3.) Hofer's work has subsequently bubbled off a large industry in contact dynamics, and we mention only a few highlights. The properties of holomorphic curves in symplectizations were analyzed by Hofer–Wysocki–Zehnder [39], [40], [41]. Also Eliashberg, Givental and Hofer [17] have suggested a *Symplectic Field Theory*, a Floer-type theory involving closed orbits of Reeb vector fields and holomorphic curves "bounding" these closed orbits. The technical details of this theory have finally started to appear – see [42].

The Weinstein conjecture in dimension three has been verified for contact structures which admit planar open book decompositions [1] (also see related work of Etnyre [21]), for certain Stein fillable contact structures [5], [71], and for certain universally tight contact structures on toroidal manifolds [3]. We also refer the reader to the survey article by Hofer [38].

In [9], Colin and the author prove the following:

Theorem 4.2 (Colin–Honda [9]). *The Weinstein conjecture holds for contact structures (M, ξ) which have open books (S, h) with periodic monodromy.* (*Here S may have many boundary components.*)

Suppose ∂S is connected. If the fractional Dehn twist coefficient $c < 0$, then (M, ξ) is overtwisted by Theorem 3.3 above. If $c = 0$, then M is a connected sum, and if $0 < c < \frac{1}{2g(S)-1}$, then the universal cover of M is S^3. Here $g(S)$ is the genus of the closed surface obtained by capping off S with a disk. In all the above cases, the Weinstein conjecture has been settled by Hofer [37].

It remains to examine the case where $c \geq \frac{1}{2g(S)-1}$. The following is the main result in [9]:

Theorem 4.3 (Colin–Honda [9]). *The cylindrical contact homology of (M, ξ), as defined below, exists and is nonzero if $c \geq \frac{1}{2g(S)-1}$.*

The nontriviality of the cylindrical contact homology for (M, ξ) implies the Weinstein conjecture for (M, ξ).

Contact homology is the simplest version of Symplectic Field Theory which takes place in the symplectization $(\mathbb{R} \times M, d(e^t\alpha))$ and counts punctured J-holomorphic spheres $\tilde{u}\colon \Sigma \to \mathbb{R} \times M$ with one positive end and many negative ends. (Here t is the coordinate for $\mathbb{R}$.) *Cylindrical* contact homology, if it exists, is a version which only counts J-holomorphic cylinders, i.e., spheres with two punctures. The cylindrical theory exists, if the following condition holds:

Condition 4.4. There exists a nondegenerate Reeb vector field R for which no contractible periodic orbit γ with Conley–Zehnder index $\mu(\gamma) = 0$, 1 or 2 bounds a finite energy plane (at the positive end) in the symplectization $\mathbb{R} \times M$.

Assuming Condition 4.4, we now define the *cylindrical contact homology*.

Let α be a contact 1-form for which R is a nondegenerate Reeb vector field, and let J be an almost complex structure on $\mathbb{R}\times M$ which is *adapted* to the symplectization: If we write $T_{(t,x)}(\mathbb{R}\times M) = \mathbb{R}\frac{\partial}{\partial t}\oplus\mathbb{R}R\oplus\xi$, then J maps ξ to itself and sends $\frac{\partial}{\partial t}\mapsto R$, $R\mapsto -\frac{\partial}{\partial t}$. Let $\mathcal{P}$ be the collection of closed orbits of R. (We may need to omit certain closed orbits, but we will not worry about these technicalities here.)

If γ is a contractible periodic orbit which bounds a disk D, then we trivialize $\xi|_D$ and define the *Conley–Zehnder index* $\mu(\gamma, D)$ to be the Conley–Zehnder index of the path of symplectic maps $\{d\phi_t : \xi_{\gamma(0)} \to \xi_{\gamma(t)}, t\in[0,T]\}$ with respect to this trivialization, where ϕ_t is the time t flow of the Reeb vector field R and T is the period of γ. In our case, M is Seifert fibered and $\pi_2(M)=0$, so $\mu(\gamma)$ is independent of the choice of D. If γ, γ' are not contractible, but belong to the same *free homotopy class* $[\gamma]=[\gamma']$, then let Z be the cylinder between γ and γ'. Trivialize $\xi|_Z$ and define the *relative* Conley–Zehnder index $\mu(\gamma,\gamma')$ to be the Conley–Zehnder index of γ minus the Conley–Zehnder index of γ', both calculated with respect to this trivialization. Again, in our case, $\mu(\gamma,\gamma')$ does not depend on our choice of Z.

Define the moduli space

$$\mathcal{M}(J,\gamma_+,\gamma_-) = \left\{ \begin{array}{c} J\text{-holomorphic cylinders } \tilde{u}=(a,u)\colon \mathbb{R}\times S^1\to\mathbb{R}\times M, \\ \lim_{s\to\pm\infty} u(s,t)=\gamma_\pm(t), \quad \lim_{s\to\pm\infty} a(s,t)=\pm\infty \end{array} \right\}.$$

Here, $\gamma_\pm(t)$ refers to some parametrization of the trajectory $\gamma_\pm$. The convergence for $u(s,t)$ and $a(s,t)$ is in the C^0-topology. The complex structure j on $\mathbb{R}\times S^1$ is the usual one: if (s,t) are coordinates on $\mathbb{R}\times\mathbb{R}/\mathbb{Z}$, then $j\colon \frac{\partial}{\partial s}\mapsto\frac{\partial}{\partial t}$. We choose a *regular* J (still adapted to the symplectization) for which $\mathcal{M}(J,\gamma_+,\gamma_-)$ is a transverse zero set of the $\overline{\partial}$-operator and has the expected dimension $\mu(\gamma_+,\gamma_-)$.

The chain group is the $\mathbb{Q}$-vector space $C=\mathbb{Q}\langle\mathcal{P}\rangle$ generated by $\mathcal{P}$. Now the boundary map $\partial\colon C\to C$ is given on elements $\gamma\in\mathcal{P}$ by:

$$\partial\gamma = \sum_{\gamma'\in\mathcal{P},\ \mu(\gamma,\gamma')=1} \frac{n_{\gamma,\gamma'}}{\kappa(\gamma')}\ \gamma',$$

where $\kappa(\gamma)$ is the multiplicity of γ. If $\mu(\gamma,\gamma')=1$, then $\mathcal{M}(\gamma,\gamma')$ is a 1-dimensional moduli space and we quotient out by translations in the $\mathbb{R}$-direction. Then $n_{\gamma,\gamma'}$ is a signed count of points in $\mathcal{M}(\gamma,\gamma')/\mathbb{R}$, following a coherent orientation scheme given in [17]. If γ, γ' are multiply covered, then each non-multiply-covered holomorphic curve $\tilde{u}\in\mathcal{M}(\gamma,\gamma')/\mathbb{R}$ contributes $\pm\kappa(\gamma)\kappa(\gamma')$ to $n_{\gamma,\gamma'}$. If $\tilde{u}$ is a k-fold cover of a somewhere injective holomorphic curve, then it is counted as $\pm k(\kappa(\gamma)\kappa(\gamma'))$. The definition of ∂ is extended linearly to all of C. (For the purposes of Theorem 4.3, we can restrict attention to the portion of $\mathcal{P}$ consisting of non-multiply-covered orbits, so we may work with $\mathbb{Z}/2\mathbb{Z}$-coefficients.)

With the above restriction (Condition 4.4) on the Conley–Zehnder indices of contractible periodic orbits, it can be shown that $\partial\circ\partial=0$, hence the cylindrical contact

homology is well-defined. Moreover, it does not depend on the choice of generic J or on the choice of nondegenerate R, provided Condition 4.4 is satisfied.

We now indicate some elements of the proof of Theorem 4.3. The well-definition of cylindrical contact homology is proved by projecting a finite energy plane $\tilde{u} = (a, u) \colon \mathbb{R}^2 \to \mathbb{R} \times M$ to M, observing that $u \colon \mathbb{R}^2 \to M$ is positively transverse to the Reeb vector field R except at complex branch points, and using (a generalization of) the Rademacher function. A similar technique gives restrictions on pairs γ, γ' which admit holomorphic cylinders between them. Since we have enough restrictions on the boundary maps, an Euler characteristic argument gives the result.

In the pseudo-Anosov case we expect the analog of Theorem 4.3 to hold when $c > \frac{1}{n}$, where n is the number of attracting (= number of repelling) periodic points. (These are currently being worked out in [10].) It still remains to consider $c = \frac{1}{n}$ in the pseudo-Anosov case….

Acknowledgements. The author wholeheartedly thanks Francis Bonahon for many suggestions on improving the exposition.

References

[1] Abbas, C., Cieliebak, K., and Hofer, H., The Weinstein conjecture for planar contact structures in dimension three. *Comment. Math. Helv.* **80** (2005), 771–793.

[2] Bennequin, D., Entrelacements et équations de Pfaff. In *Third Schnepfenried geometry conference* (Schnepfenried, 1982), Vol. 1, *Astérisque* **107–108** (1983), 87–161.

[3] Bourgeois, F., and Colin, V., Homologie de contact des variétés toroïdales. *Geom. Topol.* **9** (2005) 299–313 (electronic).

[4] Calegari, D., and Dunfield, N., Laminations and groups of homeomorphisms of the circle. *Invent. Math.* **152** (2003), 149–204.

[5] Chen, W., Pseudo-holomorphic curves and the Weinstein conjecture. *Comm. Anal. Geom.* **8** (2000), 115–131.

[6] Colin, V., Une infinité de structures de contact tendues sur les variétés toroïdales. *Comment. Math. Helv.* **76** (2001), 353–372.

[7] Colin, V., Giroux, E., and Honda, K., On the coarse classification of tight contact structures. In *Topology and geometry of manifolds* (Athens, GA, 2001), Proc. Sympos. Pure Math. 71, Amer. Math. Soc., Providence, RI, 2003, 109–120.

[8] Colin, V., Giroux, E., and Honda, K., Finitude homotopique et isotopique des structures de contact tendues. In preparation.

[9] Colin, V., and Honda, K., Reeb vector fields and open book decompositions I: the periodic case. Preprint, 2006.

[10] Colin, V., and Honda, K., Reeb vector fields and open book decompositions II: the pseudo-Anosov case. In preparation.

[11] Ding F., and Geiges, H., Symplectic fillability of tight contact structures on torus bundles. *Algebr. Geom. Topol.* **1** (2001), 153–172 (electronic).

[12] Eliashberg, Y., Contact 3-manifolds twenty years since J. Martinet's work. *Ann. Inst. Fourier* (*Grenoble*) **42** (1992), 165–192.

[13] Eliashberg, Y., Classification of overtwisted contact structures on 3-manifolds. *Invent. Math.* **98** (1989), 623–637.

[14] Eliashberg, Y., Filling by holomorphic discs and its applications. In *Geometry of low-dimensional manifolds.* 2 (Durham, 1989), London Math. Soc. Lecture Note Ser. 151, Cambridge University Press, Cambridge 1990, 45–67.

[15] Eliashberg, Y., Unique holomorphically fillable contact structure on the 3-torus. *Internat. Math. Res. Notices* **1996** (1996), 77–82.

[16] Eliashberg, Y., A few remarks about symplectic filling. *Geom. Topol.* **8** (2004), 277–293 (electronic).

[17] Eliashberg, Y., Givental, A., and Hofer, H., Introduction to symplectic field theory. *GAFA 2000* (Tel Aviv, 1999), *Geom. Funct. Anal.* (2000), Special Volume, Part II, 560–673.

[18] Eliashberg, Y., and Gromov, M., Convex symplectic manifolds. In *Several complex variables and complex geometry*, (Santa Cruz, CA, 1989), Part 2, Proc. Sympos. Pure Math. 52, Part 2, Amer. Math. Soc., Providence, RI, 1991, 135–162.

[19] Eliashberg, Y., and Thurston, W., *Confoliations*. University Lecture Series 13, Amer. Math. Soc., Providence, RI, 1998.

[20] Etnyre, J., On symplectic fillings. *Algebr. Geom. Topol.* **4** (2004), 73–80 (electronic).

[21] Etnyre, J., Planar open book decompositions and contact structures. *Internat. Math. Res. Notices* **2004** (2004), 4255–4267.

[22] Etnyre, J., and Honda, K., On the nonexistence of tight contact structures. *Ann. of Math.* (2) **153** (2001), 749–766.

[23] Etnyre, J., and Honda, K., Tight contact structures with no symplectic fillings. *Invent. Math.* **148** (2002), 609–626.

[24] Feehan, P., and Leness, T., On Donaldson and Seiberg-Witten invariants. In *Topology and geometry of manifolds* (Athens, GA, 2001), Proc. Sympos. Pure Math. 71, Amer. Math. Soc., Providence, RI, 2003, 237–248.

[25] Fenley, S., Laminar free hyperbolic 3-manifolds. Preprint, 2002, ArXiv:math.GT/0210482.

[26] Floer, A., Instanton homology and Dehn surgery. In *The Floer memorial volume*, Progr. Math. 133, Birkhäuser, Basel 1995, 77–97.

[27] Gabai, D., Foliations and the topology of 3-manifolds. II. *J. Differential Geom.* **26** (1987), 461–478.

[28] Gabai, D., and Oertel, U., Essential laminations in 3-manifolds. *Ann. of Math.* (2) **130** (1989), 41–73.

[29] Gambaudo, J.-M., and Ghys, E., Braids and signatures. *Bull. Soc. Math. France*, to appear.

[30] Ghiggini, P., Strongly fillable contact 3-manifolds without Stein fillings. *Geom. Topol.* **9** (2005), 1677–1687 (electronic).

[31] Ghiggini, P., Infinitely many universally tight contact manifolds with trivial Ozsváth-Szábó contact invariants. Preprint, 2005, ArXiv:math.GT/0510574.

[32] Giroux, E., Convexité en topologie de contact. *Comment. Math. Helv.* **66** (1991), 637–677.

[33] Giroux, E., Géométrie de contact: de la dimension trois vers les dimensions supérieures. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 405–414.

[34] Gompf, R., Handlebody construction of Stein surfaces. *Ann. of Math.* (2) **148** (1998), 619–693.

[35] Goodman, N., Overtwisted open books from sobering arcs. *Algebr. Geom. Topol.* **5** (2005), 1173–1195 (electronic).

[36] Gromov, M., Pseudo-holomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), 307–347.

[37] Hofer, H., Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three. *Invent. Math.* **114** (1993), 515–563.

[38] Hofer, H., Dynamics, topology, and holomorphic curves. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. I, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 255–280 (electronic).

[39] Hofer, H., Wysocki, K., and Zehnder, E., Properties of pseudo-holomorphic curves in symplectizations I: asymptotics. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **13** (1996), 337–379.

[40] Hofer, H., Wysocki, K., and Zehnder, E., Properties of pseudo-holomorphic curves in symplectisations II: embedding controls and algebraic invariants. *Geom. Funct. Anal.* **5** (1995), 270–328.

[41] Hofer, H., Wysocki, K., and Zehnder, E., Properties of pseudoholomorphic curves in symplectizations. III. Fredholm theory. In *Topics in nonlinear analysis*, Progr. Nonlinear Differential Equations Appl. 35, Birkhäuser, Basel 1999, 381–475.

[42] Hofer, H., Wysocki, K., and Zehnder, E., Polyfolds and Fredholm theory, Part I. Preprint, 2005.

[43] Honda, K., On the classification of tight contact structures I. *Geom. Topol.* **4** (2000), 309–368. (electronic).

[44] Honda, K., Kazez, W., and Matić, G., Convex decomposition theory. *Internat. Math. Res. Notices* **2002** (2002), 55–88.

[45] Honda, K., Kazez, W., and Matić, G., Right-veering diffeomorphisms of compact surfaces with boundary I. Preprint, 2005.

[46] Honda, K., Kazez, W., and Matić, G., Right-veering diffeomorphisms of compact surfaces with boundary II. Preprint, 2006.

[47] Honda, K., Kazez, W., and Matić, G. In preparation.

[48] Kronheimer, P., and Mrowka, T., Monopoles and contact structures. *Invent. Math.* **130** (1997), 209–255.

[49] Kronheimer, P., and Mrowka, T., Witten's conjecture and property P. *Geom. Topol.* **8** (2004), 295–310 (electronic).

[50] Kronheimer, P., Mrowka, T., Ozsváth, P., and Szabó, Z., Monopoles and lens space surgeries. *Ann. of Math.*, to appear.

[51] Lisca, P., and Stipsicz, A., An infinite family of tight, not semi-fillable contact three-manifolds. *Geom. Topol.* **7** (2003), 1055–1073 (electronic).

[52] Lisca, P., and Stipsicz, A., Tight, not semi-fillable contact circle bundles. *Math. Ann.* **328** (2004), 285–298.

[53] Lisca, P., and Stipsicz, A., Ozsváth-Szabó invariants and tight contact three-manifolds. I. *Geom. Topol.* **8** (2004), 925–945 (electronic).

[54] Lisca, P., and Stipsicz, A., Ozsváth-Szabó invariants and tight contact three-manifolds, II. Preprint, 2004, ArXiv:math.SG/0404136.

[55] Loi, A., and Piergallini, R., Compact Stein surfaces with boundary as branched covers of B^4. *Invent. Math.* **143** (2001), 325–348.

[56] Nielsen, J., Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen I. *Acta Math.* **50** (1927), 189–358.

[57] Nielsen, J., Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen II. *Acta Math.* **53** (1929), 1–76.

[58] Nielsen, J., Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen III. *Acta Math.* **58** (1931), 87–167.

[59] Ozsváth, P., and Szabó, Z., Holomorphic disks and topological invariants for closed three-manifolds. *Ann. of Math.* (2) **159** (2004), 1027–1158.

[60] Ozsváth, P., and Szabó, Z., Holomorphic disks and three-manifold invariants: properties and applications. *Ann. of Math.* (2) **159** (2004), 1159–1245.

[61] Ozsváth, P., and Szabó, Z., Heegaard Floer homology and contact structures. *Duke Math. J.* 129 (2005), 39–61.

[62] Plamenevskaya, O., Contact structures with distinct Heegaard Floer invariants. *Math. Res. Lett.* **11** (2004), 547–561.

[63] Plamenevskaya, O., Transverse knots, branched double covers and Heegaard Floer contact invariants. Preprint, 2004, ArXiv:math.GT/0412183.

[64] Roberts, R., Taut foliations in punctured surface bundles, I. *Proc. London Math. Soc.* (3) **82** (2001), 747–768.

[65] Roberts, R., Taut foliations in punctured surface bundles, II. *Proc. London Math. Soc.* (3) **83** (2001), 443–471.

[66] Roberts, R., Shareshian, J., and Stein, M., Infinitely many hyperbolic 3-manifolds which contain no Reebless foliation. *J. Amer. Math. Soc.* **16** (2003), 639–679.

[67] Taubes, C., The Seiberg-Witten invariants and symplectic forms. *Math. Res. Lett.* **1** (1994), 809–822.

[68] Thurston, W., On the geometry and dynamics of diffeomorphisms of surfaces. *Bull. Amer. Math. Soc.* **19** (1988), 417–431.

[69] Thurston, W., and Winkelnkemper, H., On the existence of contact forms. *Proc. Amer. Math. Soc.* **52** (1975), 345– 347.

[70] Torisu, I., Convex contact structures and fibered links in 3-manifolds. *Internat. Math. Res. Notices* **2000** (2000), 441–454.

[71] Zehmisch, K., Strong fillability and the Weinstein conjecture, ArXiv:math.SG/0405203.

Department of Mathematics, University of Southern California, Los Angeles, CA 90089, U.S.A.

E-mail: khonda@math.usc.edu

Generalized triangle inequalities and their applications

Michael Kapovich*

Abstract. We present a survey of our recent work on generalized triangle inequalities in infinitesimal symmetric spaces, nonpositively curved symmetric spaces and Euclidean buildings. We also explain how these results can be used to analyze some basic problems of algebraic group theory including the problem of decomposition of tensor products of irreducible representations of complex reductive Lie groups. Among the applications is a generalization of the Saturation Theorem of Knutson and Tao to Lie groups other than $\mathrm{SL}(n, \mathbb{C})$.

Mathematics Subject Classification (2000). Primary 22E46, 20G15, 53C20; Secondary 14L24, 20E42.

Keywords. Symmetric spaces, buildings, tensor product decomposition.

1. Introduction

As we learn in school, given 3 positive numbers a, b, c satisfying the familiar triangle inequalities $a \leq b + c$, etc., one can construct a triangle in the Euclidean plane whose side-lengths are a, b and c. A brief contemplation shows that the same elementary geometry proof works in the hyperbolic plane and, more generally, in all simply-connected complete nonpositively curved Riemannian manifolds.

At the first glance, it appears that this is all one can say about the triangle inequalities. Note however that in all negatively curved simply-connected symmetric spaces, the *metric length* of a geodesic segment is a complete congruence invariant. On the other hand, in higher rank symmetric spaces, the congruence classes of oriented segments are parameterized by the Weyl chamber Δ. We will refer to the parameter $\sigma(\gamma) \in \Delta$ corresponding to an oriented segment γ as its Δ-*length*. The same notion of Δ-length can be defined in Euclidean buildings, where Δ is the Weyl chamber for the finite Weyl group in the associated Euclidean Coxeter complex. In this survey we discuss our recent work appearing in a series of papers with Bernhard Leeb and John Millson ([KLM1], [KLM2], [KLM3], [KM1], [KM2]). It originates with the following basic question:

Question 1.1. Suppose that X is a nonpositively curved simply-connected symmetric space or a Euclidean building. What restrictions on the triples $(\lambda, \mu, \nu) \in \Delta^3$ are

*During the writing of this paper the author was partially supported by the NSF grant DMS-04-05180.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

necessary and sufficient for existence of an oriented geodesic triangle in X whose Δ-side lengths are λ, μ, ν?

We will see how this question (and related problems) connects to the theory of algebraic groups over real, complex and nonarchimedean valued fields as well as to representation theory of complex reductive Lie groups.

Acknowledgments. I am grateful to all my collaborators for the joint work. I am especially grateful to John Millson for our continuous collaboration in the last 13 years.

2. Metric spaces modelled on Coxeter complexes

Let A be a (finite-dimensional) Euclidean space and W_{aff} be a group of isometries of A generated by reflections in a family of hyperplanes $H \subset A$ (called *walls*). A *half-apartment* is a closed half-space in A bounded by a wall. By choosing the origin $o \in A$ and taking linear parts of the elements of W_{aff} we obtain a group $W = W_{\mathrm{sph}}$ fixing an origin o in A. We require W to be finite. Then W is a finite Coxeter group and the group W_{aff} is called an affine Coxeter (or Weyl) group. The pair (A, W_{aff}) is called a *Euclidean Coxeter complex*. We let $\Delta \subset A$ denote a fundamental domain of the reflection group W: it is a convex cone in A with vertex at o. A point $x \in A$ is called *special* if its stabilizer in W_{aff} is isomorphic to W. It turns out that each Euclidean Coxeter complex has a special point; in what follows we will always assume that $o \in A$ is special.

Example 2.1. Suppose that W is a finite Coxeter group and $W_{\mathrm{aff}} = W \ltimes V$, where V is the full group of translations of A. Such affine Coxeter groups appear naturally in the context of symmetric spaces. Another useful example to keep in mind is given by $W \ltimes Q(R^\vee)$, where W is the finite Weyl group associated with a root system $R \subset V^*$ and $Q(R^\vee) \subset V$ is the coroot lattice of R. In this case W_{aff} is discrete. Such examples appear in the context of Bruhat–Tits buildings associated with groups $\underline{G}(\mathbb{K})$, where $\underline{G}$ is a reductive algebraic group and $\mathbb{K}$ is a field with discrete valuation.

Let Z be a metric space. A *geometric structure* on Z *modelled on* the Euclidean Coxeter complex (A, W_{aff}) consists of an atlas of isometric embeddings $\varphi\colon A \hookrightarrow Z$ satisfying the following compatibility condition:

> For any two charts φ_1 and φ_2, the transition map $\varphi_2^{-1} \circ \varphi_1$ is the restriction of an element of W_{aff}.

The charts and their images, $\varphi(A) \subset Z$, are called *apartments*. We will require that any two points in Z lie in a common apartment. All W_{aff}-invariant notions introduced for the Coxeter complex (A, W_{aff}), such as walls, special points, etc., carry over to geometries modelled on (A, W_{aff}).

Thickness of a space X modelled on (A, W_{aff}) is the cardinality of the set of half-apartments adjacent to a wall in X. In all examples considered in this survey, thickness will be independent of the wall in X. The space X is called *thick* if it has thickness ≥ 3.

Examples of geometric structures modelled on (A, W_{aff}) are provided by simply-connected symmetric spaces of nonpositive curvature (in which case W_{aff} acts transitively on A), Euclidean buildings and *infinitesimal symmetric spaces* $\mathfrak{p}$. The latter is equal to the tangent space to a symmetric space, $\mathfrak{p} = T_o X$. Apartments in $\mathfrak{p}$ correspond to Cartan subalgebras (i.e., the maximal abelian subalgebras) in $\mathfrak{p}$, where $\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}$ is the Cartan decomposition of the Lie algebra $\mathfrak{g}$. Although, as a metric space, $\mathfrak{p}$ is nothing but a Euclidean space, its natural group of automorphisms is smaller than $\text{Isom}(\mathfrak{p})$, it is the *Cartan motion group* $K \ltimes \mathfrak{p}$.

Remark 2.2. *Discrete* Euclidean buildings (i.e. the ones with the discrete structure group W_{aff}) can be thought of as both geometric and combinatorial objects. From the combinatorial standpoint, one regards buildings as *polysimplical complexes*. The distance between cells is then a certain W_{aff}-valued function, paths in buildings are replaced by *galleries*, etc. Geometric viewpoint appears more powerful as far as the problems raised in this paper are concerned. For one thing, one can do analysis (rather than combinatorics) on such spaces. As another example, one can *stretch* a piecewise-linear path in an apartment via a homothety while there is no obvious stretching construction for galleries. Importance of stretching will become apparent in Section 5.3.

An important example of a symmetric space to keep in mind is Sym_n, the space of positive-definite symmetric $n \times n$ matrices with real coefficients. Then $\text{Sym}_n = \text{GL}(n, \mathbb{R})/\text{O}(n)$ and the Weyl group of this space is the permutation group S_n.

Similarly, one defines metric spaces modelled on spherical Coxeter complexes. The most important examples of such spaces are spherical buildings.

For a metric space Z modelled on (A, W_{aff}), we define the Δ-valued distance function

$$d_\Delta \colon Z \times Z \to \Delta$$

as follows:

> Given points $x, y \in Z$, find an apartment $\varphi \colon A \to A' \subset Z$ whose image contains x and y. Then consider the vector $\varphi^{-1}(v)$ in A with the tip y and tail x and project this vector to the Weyl chamber Δ via the quotient map $A \to A/W = \Delta$.

Clearly, this definition is independent of the choice of an apartment $\varphi(A)$ containing x, y.

Example 2.3. If $Z = \text{Sym}_n$, then $d_\Delta(x, y)$ is the set of eigenvalues of the matrix $x^{-1}y$, counted with multiplicity and arranged in the decreasing order.

Given the notion of Δ-valued distance between points in Z we can also define the Δ-*length* for piecewise-geodesic paths p in Z by taking the sum of the Δ-lengths of the geodesic subsegments of p.

Observe that the Δ-distance function d_Δ is not (in general) symmetric, however

$$d_\Delta(x, y) = (d_\Delta(y, x))^*,$$

where the vector $v^* = w_0(-v)$ is contragredient to the vector v. (Here w_0 is the longest element of W.)

It follows from the Cartan decomposition that in the case when Z is a (nonpositively curved) symmetric space or an infinitesimal symmetric space, then d_Δ is a complete congruence invariant of an oriented geodesic segment $\overline{xy} \subset Z$:

> There exists an automorphism $g \in \mathrm{Aut}(Z)$ which carries $\overline{x_1y_1}$ to $\overline{x_2y_2}$ if and only if $d_\Delta(x_1, y_1) = d_\Delta(x_2, y_2)$.

The situation in the case of Euclidean buildings is more subtle, we will return to this in Section 4.

Definition 2.4. Given a space X modelled on an affine Coxeter complex, we let $D_n(X)$ denote the collection of tuples $(\lambda_1, \dots, \lambda_n) \in \Delta^n$ such that there exists an oriented geodesic polygon in X with the Δ-side lengths $\lambda_1, \dots, \lambda_n$.

Thus Question 1.1 in the introduction is asking for a description of $D_3(X)$ for the given space X.

3. Generalized triangle inequalities

Suppose that X is either a nonpositively curved simply-connected symmetric space, an infinitesimal symmetric space or a thick Euclidean building, modelled on (A, W_{aff}). A priori, $D_n(X)$ is just a subset in Δ^n. The following theorem establishes basic structural properties of this set.

Theorem 3.1 ([KLM1], [KLM2]). 1. *$D_n(X)$ is a convex homogeneous polyhedral cone.*

2. *$D_n(X)$ depends only on the pair (A, W) and nothing else, not even the type of the space X (i.e., whether this is an infinitesimal symmetric space, symmetric space or a building).*

Corollary 3.2. 1. *If $\mathfrak{p} = T_oX$, where X is a nonpositively curved symmetric space, then $D_n(\mathfrak{p}) = D_n(X)$.*

2. *Suppose that $\underline{G}$ is a split reductive algebraic group, $G_1 = \underline{G}(\mathbb{C})$, $G_2 = \underline{G}(\mathbb{R})$ and $K_i \subset G_i$ are maximal compact subgroups, $i = 1, 2$. Then*

$$D_n(G_1/K_1) = D_n(G_2/K_2).$$

Since $D_n(X)$ is a convex homogeneous cone, it is defined by a system of homogeneous linear inequalities which we will refer to as *generalized triangle inequalities*. Theorem 3.1 reduces the computation of these inequalities to the case of symmetric spaces. Since, clearly, $D_n(X \times \mathbb{R}^m) = D_n(X) \times \mathbb{R}^m$, it suffices to consider spaces $X = G/K$, so that the identity component G of $\mathrm{Isom}(X)$ is semisimple. One of the main results of [KLM1] is a description of $D_n(X)$ in terms of the Schubert calculus in the Grassmannians associated to complex and real Lie groups G (i.e., the quotients G/P where P is a maximal parabolic subgroup of G).

The *Tits boundary* $\partial_{\mathrm{Tits}} X$ of X is a spherical building modelled on a spherical Coxeter complex (S, W) with spherical Weyl chamber $\Delta_{\mathrm{sph}} \subset S$. It is formed by equivalence classes of geodesic rays in X; the metric on $\partial_{\mathrm{Tits}} X$ is given by the *Tits angle* $\angle_{\mathrm{Tits}}$, see for instance [Ba]. We identify S with an apartment in $\partial_{\mathrm{Tits}} X$. Let Δ denote the Weyl chamber of X. We identify Δ_{sph} with $\partial_{\mathrm{Tits}} \Delta$.

Let B be the stabilizer of Δ_{sph} in G. For each vertex ζ of $\partial_{\mathrm{Tits}} X$ one defines the generalized Grassmannian $\mathrm{Grass}_\zeta = G\zeta = G/P$. (Here P is the maximal parabolic subgroup of G stabilizing ζ.) It is a compact homogeneous space stratified into B-orbits called *Schubert cells*. Every Schubert cell is of the form $C_\eta = B\eta$ for a unique vertex $\eta \in W\zeta \subset S^{(0)}$ of the spherical Coxeter complex. The closures $\overline{C}_\eta$ are called *Schubert cycles*. They are unions of Schubert cells and represent well defined elements in the homology $H_*(\mathrm{Grass}_\zeta, \mathbb{Z}_2)$.

For each vertex ζ of Δ_{sph} and each n-tuple $\vec{\eta} = (\eta_1, \ldots, \eta_n)$ of vertices in $W\zeta$ consider the following homogeneous linear inequality for $\xi \in \Delta^n$:

$$\sum_i \xi_i \cdot \eta_i \leq 0 \qquad (*_{\zeta;\vec{\eta}})$$

Here we identify the η_i's with unit vectors in Δ.

Let $I_{\mathbb{Z}_2}(G)$ be the set consisting of all data $(\zeta, \vec{\eta})$ such that the intersection of the Schubert classes $[\overline{C}_{\eta_1}], \ldots, [\overline{C}_{\eta_n}]$ in $H_*(\mathrm{Grass}_\zeta, \mathbb{Z}_2)$ equals $[pt]$.

Theorem 3.3 ([KLM1]). *$D_n(X) \subset \Delta^n$ consists of all solutions ξ to the system of inequalities $(*_{\zeta;\vec{\eta}})$ where $(\zeta, \vec{\eta})$ runs through $I_{\mathbb{Z}_2}(G)$.*

Remark 3.4. This system of inequalities depends on the Schubert calculus for the generalized Grassmannians G/P associated to the group G.

Typically, the system of inequalities in Theorem 3.3 is redundant. If G is a *complex* Lie group one can use the complex structure to obtain a smaller system of inequalities. In this case, the homogeneous spaces Grass_ζ are complex manifolds and the Schubert cycles are complex subvarieties and hence represent classes in *integral* homology. Let $I_{\mathbb{Z}}(G) \subset I_{\mathbb{Z}_2}(G)$ be the subset consisting of all data $(\zeta, \vec{\eta})$ such that the intersection of the Schubert classes $[\overline{C}_{\eta_1}], \ldots, [\overline{C}_{\eta_n}]$ in $H_*(\mathrm{Grass}_\zeta, \mathbb{Z})$ equals $[pt]$.

The following analogue of Theorem 3.3 was proven independently and by completely different methods in [BS] and in [KLM1]:

Theorem 3.5 (Stability inequalities). *$D_n(X)$ consists of all solutions ξ to the system of inequalities $(*_{\zeta;\vec{\eta}})$ where $(\zeta, \vec{\eta})$ runs through $I_{\mathbb{Z}}(G)$.*

As we will see in the next section, these inequalities generalize the system of inequalities used by Klyachko in [Kly1] to solve Weyl's problem on eigenvalues of sums of Hermitian matrices.

It was proven by Knutson, Tao and Woodward [KTW] that in the case $G = \mathrm{SL}(n, \mathbb{C})$ the system of inequalities appearing in Theorem 3.5 is irredundant; on the other hand, for the root systems B_2, G_2, B_3, C_3 this system is still redundant, see [KLM1] for the rank 2 computations and [KuLM] for the rank 3 computations. P. Belkale and S. Kumar in [BK] deformed the product structure on $H_*(\mathrm{Grass}_\zeta)$ to make a smaller system of inequalities defining $D_n(X)$, which is irredundant for all root systems of rank ≤ 3. Conjecturally, the new system of inequalities is irredundant for all root systems.

4. Algebraic problems

Let $\mathbb{F}$ be either the field $\mathbb{R}$ or $\mathbb{C}$, and let $\mathbb{K}$ be a nonarchimedean valued field with discrete valuation ring $\mathcal{O}$ and the value group $\mathbb{Z}$. For simplicity, let us consider here only split reductive group $\underline{G}$ over $\mathbb{Q}$, we refer the reader to [KLM3] for the discussion in the general case. Below we consider the following four algebraic problems, labeled by *linear algebra* interpretation in the case when $\underline{G} = \mathrm{GL}(n)$. We refer the reader to Fulton's survey [Fu] for the detailed discussion of these linear algebra problems. We note here only that Problem Q1 in the case $G = \mathrm{GL}(n, \mathbb{C})$ is asking for the restrictions on eigenvalues of sums of Hermitian $n \times n$ matrices A and B, provided that the eigenvalues of A and B are given.

- **Q1. Eigenvalues of a sum.** Set $G := \underline{G}(\mathbb{F})$, let K be a maximal compact subgroup of G. Let $\mathfrak{g}$ be the Lie algebra of G, and let $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be its Cartan decomposition. Give necessary and sufficient conditions on $\lambda, \mu, \nu \in \mathfrak{p}/Ad(K)$ in order that there exist elements $A, B, C \in \mathfrak{p}$ whose projections to $\mathfrak{p}/Ad(K)$ are λ, μ and ν, respectively, so that
$$A + B + C = 0.$$

- **Q2. Singular values of a product.** Let G and K be the same as above. Give necessary and sufficient conditions on $\lambda, \mu, \nu \in K\backslash G/K$ in order that there exist elements $A, B, C \in G$ whose projections to $K\backslash G/K$ are λ, μ and ν, respectively, so that
$$ABC = 1.$$

- **Q3. Invariant factors of a product.** Set $G := \underline{G}(\mathbb{K})$ and $K := \underline{G}(\mathcal{O})$. Give necessary and sufficient conditions on $\lambda, \mu, \nu \in K\backslash G/K$ in order that there

exist elements $A, B, C \in G$ whose projections to $K\backslash G/K$ are λ, μ and ν, respectively, so that

$$ABC = 1.$$

- **Q4. Decomposing tensor products.** Let $\underline{G}^\vee$ be the Langlands dual group of $\underline{G}$. Give necessary and sufficient conditions on highest weights λ, μ, ν of irreducible representations V_λ, V_μ, V_ν of $G^\vee := \underline{G}^\vee(\mathbb{C})$ so that

$$(V_\lambda \otimes V_\mu \otimes V_\nu)^{G^\vee} \neq 0.$$

Notation 4.1. Throughout this paper we will denote by Sol(Qi, G) the set of triples (λ, μ, ν) which are solutions of Problem Qi for the group G.

Remark 4.2. Assuming that the field $\mathbb{K}$ is local (and hence its residue field is a finite field of order q) we can reformulate the condition that $(\lambda, \mu, \nu) \in$ Sol(Q3, G) as

$$m_{\lambda,\mu}(\nu^*) \neq 0,$$

where the m's are the structure constants of the spherical Hecke ring associated with the group $G = \underline{G}(\mathbb{K})$. We will discuss the geometric meaning of the constants m in the end of this section.

It turns out that the first three algebraic problems are closely related to the geometric problems discussed in the previous section. Consider for instance Problem Q2. Let $X = G/K$ be the corresponding nonpositively curved symmetric space; the group G acts on X by left multiplication, preserving the function d_Δ. Let $o \in X$ be the point stabilized by K. We identify Δ with the double coset $K\backslash G/K$.

Given elements $A, B, C \in G$ we define the polygonal chain in X with the (four) vertices

$$o, x = A(o), \quad y = AB(o), \quad z = ABC(o).$$

Since G preserves the Δ-distance, we conclude that

$$d_\Delta(o, x) = \lambda, \quad d_\Delta(x, y) = \mu, \quad d_\Delta(y, z) = \nu,$$

where A, B, C project to the vectors λ, μ, ν in Δ. If $ABC = 1$, the polygonal chain yields a geodesic triangle in X; conversely, if $z = o$ then, by multiplying C by an element of K if necessary, we get $ABC = 1$. Therefore

$$D_3(X) = \text{Sol(Q2, } G), \quad \text{for the symmetric space } X = G/K.$$

The same arguments work for the infinitesimal symmetric space $X' = T_oX$:

$$D_3(X') = \text{Sol(Q1, } G).$$

The situation in the case of Problem Q3 is more subtle. It is easy to see that

$$\text{Sol(Q3, } G) \subset D_3(X),$$

where X is the (discrete) Bruhat–Tits building corresponding to the group G. There are two straightforward restrictions on the elements of Sol(Q3, G) $\subset D_3(X)$:

- **O1.** $\mathrm{Sol}(\mathrm{Q3}, G) \subset L^3$, where L is the cocharacter lattice of a maximal torus in $\underline{G}$.
- **O2.** For each triple $\sigma = (\lambda, \mu, \nu) \in \mathrm{Sol}(\mathrm{Q3}, G)$ we have

$$T(\sigma) := \lambda + \mu + \nu \in Q(R^\vee).$$

We let $\Lambda \subset L^3$ denote the set of triples σ satisfying $T(\sigma) \in Q(R^\vee)$. Then

Theorem 4.3 ([KLM3]). 1. $(\lambda, \mu, \nu) \in \mathrm{Sol}(\mathrm{Q3}, G)$ *if and only if there exists a geodesic triangle* $\tau \subset X$ *whose vertices are* special points *of* X *and whose* Δ*-side lengths are* λ, μ, ν.

2. *"Conversely", if* $\sigma \in D_3(X) \cap \Lambda$ *then there exists a geodesic triangle* $\tau \subset X$ *whose vertices are* vertices *of* X *and whose* Δ*-side lengths are* λ, μ, ν.

Note that the vertices of the triangle τ in Part 2 of this theorem need not be special vertices of X, unless the root system R is of type A when *all vertices of* X *are special*. The latter is ultimately responsible for the equivalence of all 4 algebraic problems in the case of $\underline{G} = \mathrm{GL}(n)$.

The basic reason why one should not expect the equality

$$\mathrm{Sol}(\mathrm{Q3}, G) = D_3(X) \cap \Lambda$$

in general, is lack of homogeneity of Euclidean buildings X: The Δ-valued distance function is not a complete congruence invariant of pairs of points in X. One can remedy this by introducing the *refined distance function* $d_{\mathrm{ref}}(x, y)$ between points $x, y \in X$. The new distance function takes values in $A \times A/W_{\mathrm{aff}}$ (see [KLM2]).

Theorem 4.4 (Transfer Theorem, [KLM2]). *Suppose that* X, X' *are thick Euclidean buildings modelled on the same Euclidean Coxeter complex* (A, W_{aff}). *Then for each geodesic polygon* $[x_1, \ldots, x_n, x_{n+1} = x_1] \subset X$ *there exists a geodesic polygon* $[x_1', \ldots, x_n', x_{n+1}' = x_1'] \subset X'$ *so that*

$$d_{\mathrm{ref}}(x_i, x_{i+1}) = d_{\mathrm{ref}}(x_i', x_{i+1}'), \quad i = 1, \ldots, n.$$

Corollary 4.5 ([KLM3]). $\mathrm{Sol}(\mathrm{Q3}, G)$ *is independent of the field* $\mathbb{K}$. *If* $\mathbb{K}$, $\mathbb{K}'$ *are nonarchimedean valued fields with the valuation group* $\mathbb{Z}$, *then*

$$\mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K})) = \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K}')).$$

Relation to representation theory. First notice that the Langlands dual group $G^\vee = \underline{G}(\mathbb{C})$ appears naturally in the context of problems Q3 and Q4: The character lattice of a maximal torus in $G^\vee$ is the cocharacter lattice of the corresponding maximal torus in G.

The following theorem was originally proven in [KLM3] via Satake correspondence; new proofs were given by Tom Haines in [Ha1] and in the work with John Millson [KM1] (along the same lines one can give yet another proof using the results of S. Gaussent and P. Littelmann [GL]).

Theorem 4.6. $\mathrm{Sol}(\mathrm{Q4}, G^\vee) \subset \mathrm{Sol}(\mathrm{Q3}, G)$.

Thus we can summarize the above results and Theorem 3.1 as:

Theorem 4.7.

$$\begin{aligned} D_3(X) &= \mathrm{Sol}(\mathrm{Q1}, \underline{G}(\mathbb{C})) = \mathrm{Sol}(\mathrm{Q1}, \underline{G}(\mathbb{R})) = \mathrm{Sol}(\mathrm{Q2}, \underline{G}(\mathbb{C})) \\ &= \mathrm{Sol}(\mathrm{Q2}, \underline{G}(\mathbb{R})) \supset \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K})) \supset \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})), \end{aligned}$$

where $X = \underline{G}(\mathbb{C})/K$.

Remark 4.8. Different proofs of the equalities

$$\mathrm{Sol}(\mathrm{Q1}, \underline{G}(\mathbb{F})) = \mathrm{Sol}(\mathrm{Q2}, \underline{G}(\mathbb{F}))$$

were given (in case $\mathbb{F} = \mathbb{C}$) for classical groups by A. Klyachko [Kly2], for all complex Lie groups by A. Alexeev, E. Meinrenken, C. Woodward [AMW] and for all real groups by S. Evens, J.-H. Lu [EL].

To which extent can the last two inclusions in Theorem 4.7 be reversed? The restrictions O1, O2 provide necessary conditions for triples (λ, μ, ν) to belong to $\mathrm{Sol}(\mathrm{Qi}, G)$, $i = 3, 4$. The natural question is if they are also sufficient. On the negative side:

Theorem 4.9 ([KLM3]). 1. *In the case of the groups* $\mathrm{Sp}(4)$ *and* G_2 *the inclusion* $\mathrm{Sol}(\mathrm{Q2}, \underline{G}(\mathbb{R})) \cap \Lambda \supset \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K}))$ *is proper.*

2. *For all non-simply laced groups the inclusion*

$$\mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K})) \supset \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C}))$$

is proper.

It turns out however that one can reverse both inclusions at the expense of *multiplication by a saturation factor*. Let θ be the highest root of the root system R (associated with the group $\underline{G}$)). Then we have the expansion

$$\theta = \sum_{i=1}^{\ell} m_i \alpha_i,$$

where the α_i's are the simple roots in R (corresponding to the chamber Δ).

Definition 4.10. Define the *saturation factor* k_R to be the least common multiple of $m_1, \ldots, m_\ell$.

Below are the values of k_R for the irreducible root systems.

Root system R	Group G	k_R
A_ℓ	$\mathrm{SL}(\ell+1), \mathrm{GL}(\ell+1)$	1
B_ℓ	$\mathrm{SO}(2\ell+1)$	2
C_ℓ	$\mathrm{Sp}(2\ell)$	2
D_ℓ	$\mathrm{SO}(2\ell)$	2
G_2	G	6
F_4	G	12
E_6	G	6
E_7	G	12
E_8	G	60

Theorem 4.11. 1. $\mathrm{Sol}(\mathrm{Q2}, \underline{G}(\mathbb{R})) \cap k_R\Lambda \subset \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K}))$ (*see* [KLM3].)

2. $k_R \cdot \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K})) \subset \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C}))$. *Moreover, if at least one of the weights* λ, μ, ν *is a sum of minuscule dominant weights, then*

$$(\lambda, \mu, \nu) \in \mathrm{Sol}(\mathrm{Q3}, \underline{G}(\mathbb{K})) \iff (\lambda, \mu, \nu) \in \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})).$$

(*See* [KM1].)

3. *Therefore*

$$D_3(X) \cap k_R^2\Lambda \subset \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})),$$

where X *is the symmetric space of* $\underline{G}(\mathbb{F})$.

In particular, in the case when $\underline{G}$ has type A_ℓ (e.g. $\underline{G} = \mathrm{SL}(\ell+1)$) we obtain a new proof of the *Saturation Theorem* of A. Knutson and T. Tao [KT] (another proof of this theorem was later given by H. Derksen and J. Whyman in [DW]):

Theorem 4.12. *The semigroup* $\Sigma := \mathrm{Sol}(\mathrm{Q4}, \mathrm{SL}(\ell+1, \mathbb{C}))$ *is* saturated*, i.e., a triple* $\sigma = (\lambda, \mu, \nu) \in \Lambda$ *belongs to* Σ *if and only if there exists* $N \in \mathbb{N}$ *so that* $N\sigma$ *belongs to* Σ.

Proof. Since the cone $D_3(X)$ is homogeneous, the semigroup $D_3(X) \cap \Lambda$ is clearly saturated. However, according to Part 3 of Theorem 4.11, $\Sigma = D_3(X) \cap \Lambda$ in our case. □

Remark 4.13. The equality

$$\mathrm{Sol}(\mathrm{Q3}, \mathrm{GL}(\ell, \mathbb{C})) = \mathrm{Sol}(\mathrm{Q4}, \mathrm{GL}(\ell, \mathbb{C}))$$

was known since the 1960s, see [K1], [K2]. However these proofs do not generalize to other root systems.

Except for the case of the root system of type A, the constants which appear in Theorem 4.11 are (conjecturally) not optimal:

Conjecture 4.14 ([KM1], [KM2]). 1. $D_3(X) \cap k\Lambda \subset \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C}))$, where $k = 1$ in the case when the root system R is simply laced and $k = 2$ otherwise.

2. Suppose that all three dominant weights λ, μ, ν are *regular*, i.e., belong to the interior of the chamber Δ. Then

$$(\lambda, \mu, \nu) \in D_3(X) \cap \Lambda \iff (\lambda, \mu, \nu) \in \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})).$$

This conjecture holds for the rank 2 simple Lie groups (see [KM2]); it is also supported by some computational experiments.

A less ambitious form of Conjecture 4.14 is

Conjecture 4.15 (S. Kumar). If $\mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})) \neq D_3(X) \cap L^3$, then there exists a triple

$$(\lambda, \mu, \nu) \in D_3(X) \cap L^3 \setminus \mathrm{Sol}(\mathrm{Q4}, \underline{G}^\vee(\mathbb{C})),$$

so that at least one of the vectors λ, μ, ν is non-singular.

Counting triangles. Let X be a Bruhat–Tits building of thickness $q + 1 < \infty$, i.e., $q + 1$ is the number of half-apartments adjacent to each wall in X. Equivalently, q is the number of elements in the residue field of $\mathbb{K}$. Our goal is to relate the number of geodesic triangles in X with the given Δ-side lengths to the dimensions of the space of $G^\vee$–invariants

$$n_{\lambda,\mu,\nu}(0) = \dim(V_\lambda \otimes V_\mu \otimes V_\nu)^{G^\vee}.$$

Let $o \in X$ be a special point, for instance, the unique point stabilized by $\underline{G}(\mathcal{O})$. Let $f(q) := m_{\lambda,\mu,\nu}(0)$ denote the number of oriented geodesic triangles $[o, x, y]$ in X with the Δ-side lengths $\lambda, \mu, \nu \in P(R^\vee) \cap \Delta$.

Remark 4.16. The Hecke ring structure constant $m_{\lambda,\mu}(\nu^*)$ is the number of geodesic triangles as above for which the vertex y is fixed.

Given the root system R and the set R^+ of positive roots (determined by Δ), let ρ denote the half-sum of the positive roots.

Theorem 4.17 ([KLM3]). *$f(q)$ is a polynomial function of q of degree $\leq q^{\langle \rho, \lambda+\mu+\nu \rangle}$ so that*

$$f(q) = n_{\lambda,\mu,\nu}(0) q^{\langle \rho, \lambda+\mu+\nu \rangle} + \textit{ lower order terms}.$$

5. Geometry behind the proofs

5.1. Weighted Busemann functions and stability. Let X be a symmetric space of nonpositive curvature or a Euclidean building. Recall that the ideal boundary $B = \partial_{\mathrm{Tits}} X$ has the structure of a spherical building, the metric on B is denoted by $\angle_{\mathrm{Tits}}$. Given a Weyl chamber Δ in X, we get a spherical Weyl chamber $\Delta_{\mathrm{sph}} = \partial_\infty \Delta \subset \partial_{\mathrm{Tits}} X$. We will identify Δ_{sph} with the unit vectors in Δ. We have a canonical projection $\theta\colon \partial_{\mathrm{Tits}} X \to \Delta_{\mathrm{sph}}$.

Take a collection of weights $m_1, \dots, m_n \geq 0$ and define a finite measure space $(\mathbb{Z}/n\mathbb{Z}, \nu)$ where the measure ν on $\mathbb{Z}/n\mathbb{Z}$ is given by $\nu(i) = m_i$. An n-tuple of ideal points $(\xi_1, \dots, \xi_n) \in B^n$ together with $(\mathbb{Z}/n\mathbb{Z}, \nu)$ determine a *weighted configuration at infinity*, which is a map

$$\psi\colon (\mathbb{Z}/n\mathbb{Z}, \nu) \to \partial_{\mathrm{Tits}} X.$$

The *type* $\tau(\psi) = (\tau_1, \dots, \tau_n) \in \Delta^n$ of the weighted configuration ψ is given by $\tau_i = m_i \cdot \theta(\xi_i)$. Let $\mu = \psi_*(\nu)$ be the pushed forward measure on B. We define the *slope* of a measure μ on B with finite total mass $|\mu|$ as

$$\mathrm{slope}_\mu(\eta) = -\int_B \cos \angle_{\mathrm{Tits}}(\xi, \eta)\, d\mu(\xi).$$

To see where the slope function comes from, consider the *μ-weighted Busemann function* on X

$$b_\mu(x) := \int_B b_\xi(x) d\mu(\xi)$$

where $b_\xi\colon X \to \mathbb{R}$ is the Busemann function on X corresponding to the point $\xi \in \partial_{\mathrm{Tits}} X$. We normalize all Busemann functions to vanish at a certain point $o \in X$. The function b_μ is a convex $|\mu|$-Lipschitz function on X which is *asymptotically linear* along each geodesic ray $\rho = \overline{o\eta}$ in X. Then

$$\mathrm{slope}_\mu(\eta) = \lim_{t\to\infty} \frac{b_\mu(\rho(t))}{t}$$

is the *asymptotic slope* of b_μ in the direction of η.

Remark 5.1. Weighted Busemann functions are a powerful tool for studying asymptotic geometry of nonpositively curved spaces, see for instance [BCG].

In what follows we will consider only measures μ with finite support.

Definition 5.2 (Stability). A measure μ on B is called *semistable* if $\mathrm{slope}_\mu(\eta) \geq 0$ and *stable* if $\mathrm{slope}_\mu(\eta) > 0$ for all $\eta \in B$.

There is a refinement of the notion of semistability motivated by the corresponding concept in geometric invariant theory.

Definition 5.3 (Nice semistability). A measure μ on B (with finite support) is called *nice semistable* if μ is semistable and $\{\text{slope}_\mu = 0\}$ is a subbuilding or empty. In particular, stable measures are nice semistable.

A weighted configuration ψ on B is called *stable, semistable* or *nice semistable*, respectively, if the corresponding measure $\psi_*\nu$ has this property.

For our purposes, nice semistability is a useful concept in the case of symmetric spaces and infinitesimal symmetric spaces only. We note however that for these spaces, existence of a semistable configuration ψ on $\partial_{\text{Tits}}X$ implies existence of a nice semistable configuration on $\partial_{\text{Tits}}X$, which has the same type as ψ, see [KLM1].

Example 5.4. Let B be a 0-dimensional spherical building. Then a measure μ on B is stable iff it contains no atoms of mass $\geq \frac{1}{2}|\mu|$, semistable iff it contains no atoms of mass $> \frac{1}{2}|\mu|$, and nice semistable iff it is either stable or consists of two atoms of equal mass.

Suppose now that G is a reductive complex Lie group, $K \subset G$ is a maximal compact subgroup, $X = G/K$ is the associated symmetric space. Then the spaces of weighted configurations in $\partial_{\text{Tits}}X$ of the given type $\tau \in \Delta^n$ can be identified with products

$$F = F_1 \times \cdots \times F_n$$

where F_i's are smooth complex algebraic varieties (generalized flag varieties) on which the group G acts transitively. Hence G acts on F diagonally.

In case X is the symmetric space associated to a complex Lie group, the notions of stability (semistability, etc.) introduced above coincide with corresponding notions from symplectic geometry, and, in the case when τ_i's are fundamental weights, they also coincide with the concepts of stability (semistability, etc.) used in Geometric Invariant Theory, see [KLM1].

Define the subset $\Delta^n_{\text{ss}}(B) \subset \Delta^n$ consisting of those n-tuples $\tau \in \Delta^n$ for which there exists a weighted semistable configuration on B of type τ. One of the central results of [KLM1] is

Theorem 5.5. *$\Delta^n_{\text{ss}}(B)$ is a convex homogeneous cone defined by the linear inequalities $(*_{\zeta;\vec{\eta}})$.*

This theorem generalizes the results of A. Klyachko [Kly1] (in the case of $\text{GL}(n)$) and A. Berenstein and R. Sjamaar [BS] in the case of complex semisimple Lie groups.

5.2. Gauss maps and associated dynamical systems. We now relate polygons in X (where X is an infinitesimal symmetric space, a nonpositively curved symmetric space or a Euclidean building) and weighted configurations on the ideal boundary B of X, which plays a key role in [KLM1] and [KLM2].

Consider a (closed) polygon $P = [x_1, x_2, \dots x_n]$ in X, i.e., a map $\mathbb{Z}/n\mathbb{Z} \to X$. The distances $m_i = d(x_i, x_{i+1})$ determine a finite measure ν on $\mathbb{Z}/n\mathbb{Z}$ by $\nu(i) = m_i$. The polygon P gives rise to a collection Gauss(P) of *Gauss maps*

$$\psi : \mathbb{Z}/n\mathbb{Z} \to \partial_{\mathrm{Tits}} X \tag{1}$$

by assigning to i an ideal point $\xi_i \in \partial_{\mathrm{Tits}} X$ so that the geodesic ray $\overline{x_i \xi_i}$ (originating at x_i and asymptotic to ξ_i) passes through x_{i+1}.

Remark 5.6. This construction, in the case of the hyperbolic plane, appears in a letter of Gauss to Wolfgang Bolyai, [G]. I am grateful to Domingo Toledo for this observation.

Taking into account the measure ν, we view the maps $\psi : (\mathbb{Z}/n\mathbb{Z}, \nu) \to \partial_{\mathrm{Tits}} X$ as *weighted configurations* of points on $\partial_{\mathrm{Tits}} X$. Note that if X is a symmetric space and the m_i's are all non-zero, there is a unique Gauss map. On the other hand, if X is a Euclidean building then there are, in general, infinitely many Gauss maps. However, the corresponding weighted configurations are of the same type, i.e., they project to the same weighted configuration on Δ_{sph}.

The following crucial observation explains why the notion of semistability is important for studying closed polygons.

Lemma 5.7 ([KLM1], [KLM2]). *For each Gauss map ψ the push forward measure $\mu = \psi_* \nu$ is semistable. If X is a symmetric space or an infinitesimal symmetric space then the measure μ is nice semistable.*

Polygons in infinitesimal symmetric spaces X'. Let $X' = T_o X$ be the infinitesimal symmetric space. Then

Theorem 5.8 ([KLM1]). 1. *ψ is nice semistable iff the corresponding weighted Busemann function b_μ attains its minimum on X, iff b_μ has a critical point in X.*

2. *Suppose that b_μ attains its minimum at the origin $o \in X$. Identify the ideal points ξ_i with the unit vectors $\bar{\xi}_i$ in the tangent space $X' = T_o X$ via the exponential map.*

Then the gradient of b_μ at o satisfies

$$0 = \nabla_o(b_\mu) = \sum_{i=1}^{n} m_i \bar{\xi}_i.$$

Thus, in Part 2 of the above theorem, we obtain a closed polygon in the infinitesimal symmetric space X' whose Δ-side lengths are $m_i \theta(\xi_i)$.

Corollary 5.9. $\Delta^n_{\mathrm{ss}}(\partial_{\mathrm{Tits}} X) = D_n(X')$.

Polygons in nonpositively curved symmetric spaces and buildings. Our goal is to "invert Gauss maps", i.e., given a semistable weighted configuration $\psi\colon (\mathbb{Z}/n, \nu) \to B$, we would like to find a closed geodesic n-gon P so that $\psi \in \mathrm{Gauss}(P)$. The polygons P correspond to the fixed points of a certain dynamical system. For $\xi \in \partial_{\mathrm{Tits}} X$ and $t \geq 0$, define the map $\phi := \phi_{\xi,t}\colon X \to X$ by sending x to the point at distance t from x on the geodesic ray $\overline{x\xi}$. Since X is nonpositively curved, the map ϕ is 1-Lipschitz. Then, given a weighted configuration $\psi\colon (\mathbb{Z}/n\mathbb{Z}, \nu) \to \partial_{\mathrm{Tits}} X$ with non-zero total mass, define the map

$$\Phi = \Phi_\psi\colon X \to X$$

as the composition

$$\phi_{\xi_n, m_n} \circ \cdots \circ \phi_{\xi_1, m_1}.$$

The fixed points of Φ are the first vertices of closed polygons $P = [x_1, \ldots, x_n]$ so that $\psi \in \mathrm{Gauss}(P)$. Since the map Φ is 1-Lipschitz, and the space X is complete and has nonpositive curvature, the map Φ has a fixed point if and only if the dynamical system $(\Phi^i)_{i\in\mathbb{N}}$ has a bounded orbit, see [KLM2]. Of course, in general, there is no reason to expect that $(\Phi^i)_{i\in\mathbb{N}}$ has a bounded orbit: For instance, if the support of the measure $\mu = \psi_*(\nu)$ is a single point, all orbits are unbounded.

Problem 5.10. Suppose that X is a CAT(0) metric space and the weighted configuration ψ is nice semistable. Is it true that $(\Phi^i)_{i\in\mathbb{N}}$ has a bounded orbit?

Although we do not know an answer to this problem in general, we have:

Theorem 5.11. 1. *Suppose that X is a nonpositively curved simply-connected symmetric space. Then ψ is nice semistable if and only if $(\Phi^i)_{i\in\mathbb{N}}$ has a bounded orbit, see* [KLM1].

2. *Suppose that X is a Euclidean building. Then ψ is semistable if and only if $(\Phi^i)_{i\in\mathbb{N}}$ has a bounded orbit. This was proven for locally compact buildings in the original version of* [KLM2]*, for* 1*-vertex buildings in* [KLM2] *and by Andreas Balser* [B] *in the general case.*

Corollary 5.12 ([KLM1], [KLM2]). *Suppose that X is a symmetric space of nonpositive curvature or a Euclidean building. Then $D_n(X) = \Delta^n_{\mathrm{ss}}(\partial_{\mathrm{Tits}} X)$.*

Now we can explain Part 2 of Theorem 3.1. For instance, let $X_i' = T_o(X_i)$, $i = 1, 2$, be infinitesimal symmetric spaces. Then

$$D_n(X_i') = D_n(X_i') = \Delta^n_{\mathrm{ss}}(\partial_{\mathrm{Tits}} X_i), \quad i = 1, 2,$$

see Corollaries 5.9 and 5.12. Let Y_i denote the 1-vertex Euclidean building which is the Euclidean cone over the spherical building $\partial_{\mathrm{Tits}} X_i$, $i = 1, 2$. Then, $\partial_{\mathrm{Tits}} Y_i = \partial_{\mathrm{Tits}} X_i$ and hence, according to Corollary 5.9,

$$D_n(Y_i) = \Delta^n_{\mathrm{ss}}(\partial_{\mathrm{Tits}} X_i), \quad i = 1, 2.$$

Finally, according to the Transfer Theorem 4.4,

$$D_n(Y_1) = D_n(Y_2),$$

since these buildings are modelled on the same Euclidean Coxeter complex (A, W), where W is the finite Weyl group of X_i, $i = 1, 2$. By combining these equalities we obtain

$$D_n(X_1') = D_n(X_2').$$

5.3. Relation to representation theory. We now explain the connection of geodesic triangles in Euclidean buildings to the representation theory which appears in [KM1]. The key instrument here is *Littelmann's path model*. Given a thick Euclidean building X modelled on (A, W_{aff}) (and associated with a nonarchimedean Lie group $G = \underline{G}(\mathbb{K})$) and a special point $o \in X$, we define the projection

$$f\colon X \to \Delta,\ f(x) = d_\Delta(o, x).$$

This projection restricts to an isometry on each alcove $a \subset X$. Moreover, f preserves the Δ-length of piecewise-geodesic paths in X. Given a geodesic triangle $[o, x, y] \subset X$, we obtain a *broken triangle*

$$f([o, x, y]) \subset \Delta$$

which has two geodesic sides $\overline{of(x)}$, $\overline{f(y)o}$ and a broken side $p = f(\overline{xy})$. The Δ-lengths of the above paths are $\lambda = d_\Delta(o, x)$, $\nu = d_\Delta(y, o)$ and $\mu = d_\Delta(x, y)$ respectively. One of the main results of [KM1] is an intrinsic description of the piecewise-geodesic paths p which appear as the result of the above construction. They turn out to be closely related to *LS paths* introduced by Peter Littelmann in [Li].

The LS paths are defined by two axioms: The first one requires existence of a certain *chain* between the tangent vectors $p'_-(t)$, $p'_+(t)$ to the path p at each breakpoint $p(t)$. The second axiom is a maximality condition for such a chain. This axiom is vacuously true if $p(t)$ is a special point of (A, W_{aff}). In [KM1] we define *Hecke paths* p in A as piecewise-geodesic paths satisfying the 1st of Littelmann's axioms. Below are the precise definitions.

Let (A, W_{aff}) be a Euclidean Coxeter complex with V the vector space underlying A; let $W' \subset W_{\mathrm{aff}}$ be the stabilizer of a point in A. By looking at the linear parts of the elements of W', we identify W' with a subgroup

$$l(W') \subset W \subset W_{\mathrm{aff}},$$

where W is the stabilizer of the origin in W_{aff}. Let $\Delta \subset V$ be a Weyl chamber of W. Then a *W'-chain in V from $\eta_0 \in V$ to $\eta_m \in V$* is a sequence of pairwise distinct vectors

$$\eta_0, \ldots, \eta_m \in V$$

so that for each i there exists a reflection $\tau_i \in l(W')$ which sends η_i to η_{i+1} and whose fixed-point set separates η_i from Δ.

Axiom 1. A piecewise-linear path $p\colon I \to A$ is a *Hecke path* if for each $t \in I$ there is a $W_{\mathrm{aff},p(t)}$-chain from $p'_-(t)$ to $p'_+(t)$.

Here and below $W_{\mathrm{aff},x}$ is the stabilizer of x in W_{aff}.

Axiom 2. A path p satisfies the *maximality property* if for each $p(t)$ the above $W_{\mathrm{aff},x}$-chain can be found which is a *maximal W-chain* from $p'_-(t)$ to $p'_+(t)$.

Definition 5.13. A path p in A is said to be an *LS path* if it satisfies Axioms 1 and 2.

Here maximality is understood in the set-theoretic sense.

Theorem 5.14 ([KM1]). *A path p in Δ is the projection (under f) of a geodesic path in X if and only if p is a Hecke path.*

Remark 5.15. An analogous result was independently proven in [GL] by Gaussent and Littelmann in the context of *folded galleries*.

On the other hand, Littelmann proved in [Li] the following fundamental

Theorem 5.16.

$$\dim((V_\lambda \otimes V_\mu \otimes V_\nu)^{G^\vee})$$

is the number of polygons $P \subset \Delta$, each of which is the concatenation of the paths $\overline{o\lambda}$, p, $\overline{\nu^* o}$, where p is an LS path of the Δ-length μ.

Given this we immediately see that

$$\mathrm{Sol}(\mathrm{Q4}, G^\vee) \subset \mathrm{Sol}(\mathrm{Q3}, G),$$

since each LS path is also a Hecke path.

The converse relation is not as clear, since Hecke paths in general are not LS paths (maximality axiom may fail). Nevertheless, suppose for a moment that μ is an integer multiple of a fundamental coweight ϖ_i. Then the geodesic segment $\overline{xy}$ (having special end-points and Δ-length μ) crosses walls of X only at vertices of X. Therefore, each break-point $p(t)$ of the path $p = f(\overline{xy})$ is a vertex of the Coxeter complex (A, W_{aff}). One can easily see from the definition of the constant k_R (Definition 4.10) that for each vertex v of (A, W_{aff}),

$$k_R v \text{ is a special vertex of } (A, W_{\mathrm{aff}}).$$

Therefore, the rescaled path $k_R \cdot p$ satisfies the 2nd Axiom of an LS path, while the 1st Axiom is preserved by integer scalings. Hence $k_R \cdot p$ is an LS path and thus, by Littelmann's Theorem

$$(V_{k_R\lambda} \otimes V_{k_R\mu} \otimes V_{k_R\nu})^{G^\vee} \neq 0.$$

This establishes Part 2 of Theorem 4.11 provided that the coweight μ is a multiple of some fundamental coweight. The general case is more subtle, we refer the reader to [KM1] for the details.

6. Other developments

Restriction problems. The *Restriction Problem* is, in a sense, even more fundamental for the representation theory than the tensor product decomposition problem:

$$(V_\lambda \otimes V_\mu \otimes V_\nu)^G \neq 0$$

if and only if the restriction of the product representation of $G \times G \times G$ to the diagonal $G \subset G \times G \times G$ has a nonzero invariant vector.

Problem 6.1 (Restriction Problem). Let $H \subset G$ be a complex reductive subgroup in a complex reductive group G. Given an irreducible representation V_μ of G, decompose its restriction $V_\mu | H$ into irreducible factors.

Considerable progress on the general restriction problem and its infinitesimal geometric analogue (determining the projection of a cotangent orbit in the dual of the Lie algebra of a compact group to the dual of the Lie algebra of a subgroup) was made in [BS]. However it appears very difficult to prove saturation results in this generality. Below are two major obstacles in extending the results of [KLM1], [KLM2], [KLM3], [KM1] to the general restriction problem:

1. Given a subbuilding $Y \subset X$ in a Euclidean building X, there do not seem to be natural 1-Lipschitz retractions $X \to Y$. The nearest-point projection does not appear to be a good choice.

2. There is (at present) no analogue of Littelmann's path model for the general restriction problem: Littelmann's solution in [Li] of the restriction problem applies only to Levi subgroups.

It turns out, however, that the entire analysis of the *generalized triangle inequality* problems as well as the corresponding algebra problems Q1–Q4 outlined above, can be extended to the restriction problem in the case of Levi subgroups $H \subset G$. This generalization is carried out in the joint work with John Millson and Tom Haines [HKM]. Geodesic polygons are replaced with ideal geodesic polygons, for their infinite sides, the Δ-valued distance function is replaced with Δ-valued Busemann function, etc.

For instance, we obtain a generalization of the Saturation Theorem 4.11 described below.

Let G be a complex reductive Lie group, L be the character lattice of its maximal torus, $Q(R)$ be its root lattice. Let $M \subset G$ be a Levi subgroup; let $\Delta \subset \Delta_M$ be the Weyl chambers of G and M. Set

$$\mathcal{L} := \{(\lambda, \mu) \in L \times L : \lambda + \mu^* \in Q(R^\vee)\}.$$

Let k_R denote the saturation factor for the root system R of the group G, see Definition 4.10. Then there exists a convex polyhedral cone $D(M, G) \subset \Delta_M \times \Delta$ so that:

Theorem 6.2 ([HKM]). 1. *If* λ, μ *are dominant weights of* M *and* G *respectively so that*

$$V_\lambda \subset V_\mu|M,$$

then $(\lambda, \mu) \in \mathcal{L} \cap D(M, G)$.

2. *If* $(\lambda, \mu) \in \mathcal{L} \cap D(M, G)$, *then for* $k = k_R^2$,

$$V_{k\lambda} \subset V_{k\mu}|M.$$

Remark 6.3. In the case $R = A_\ell$, $k_R = 1$, and hence the above result is the analogue of the Knutson–Tao saturation theorem and in fact implies their saturation theorem.

Problem 6.4. Prove an analogue of Theorem 6.2 for arbitrary reductive subgroups H of G with constant k_R^2 replaced by a suitable number k computable in terms of G and H.

Note that even the case of $G = \mathrm{SL}(n, \mathbb{C})$ is extremely interesting in view of possible applications to P $\neq$ NP, see [MS1], [MS2].

Structure of the sets Sol(Qi). Despite the description of the convex cone Sol(Q1, G) via the linear inequalities $(*_{\zeta;\vec{\eta}})$, its structure remains somewhat mysterious. The case understood best is when $G = \mathrm{SL}(n, \mathbb{C})$.

1. For $\mathrm{SL}(n, \mathbb{C})$ there exists a procedure for computing the inequalities $(*_{\zeta;\vec{\eta}})$ by induction on the rank: This procedure was first conjectured by R. Horn in the 1960s; it was proven in a combination of works by A. Klyachko [Kly1] and A. Knutson and T. Tao [KT]. An alternative proof was later given by P. Belkale [Be2].

Problem 6.5. Generalize Horn's recursion formula to groups other than $\mathrm{SL}(n, \mathbb{C})$.

Such a generalization would give a more practical algorithm for computation of Sol(Q1, G) than the generalized Schubert calculus.

2. While the facets of the cone Sol(Q1, G) are given by the inequalities $(*_{\zeta;\vec{\eta}})$, the edges of this cone have not been described, except in the SL(n) case, see [KTW], [Be4].

3. Concerning Problems Q3 and Q4 one can reasonably ask "what a computation of the sets Sol(Q3), Sol(Q4) might mean?"

The following theorem was proven by C. Laskowski in [La] for Problem Q3 and in [KM2] for Problem Q4:

Theorem 6.6. *For each* G, *the set* Sol(Qi, G) $(i = 3, 4)$ *is a finite union of* elementary sets.

The notion of an elementary subset of L^3 comes from logic: It is a set given by a finite collection of linear inequalities (with integer coefficients) and congruences. Therefore one can interpret Problems Q3, Q4 as

Problem 6.7. Find the inequalities and congruences in the description of Sol(Q3) and Sol(Q4) as unions of elementary sets.

An example of such description is given in [KM2] for Sol(Q4, G), $G = \mathrm{Sp}(4, \mathbb{C})$, $G = G_2$. Note that for these groups, Sol($\mathbf{Qi}$, G), $i = 3, 4$, are not elementary sets themselves.

Note that if Conjecture 4.14 holds, then for each simply-laced group $\underline{G}$ the sets Sol(Q3), Sol(Q4) are elementary and are both equal to

$$D_3(X) \cap L^3,$$

where $X = \underline{G}(\mathbb{C})/K$.

Quantum product problems. Throughout this paper we restricted our attention only to noncompact Lie groups and nonpositively curved spaces. However Problem Q2 has a straightforward generalization in the case of compact Lie groups.

Let K be a maximal compact subgroup in a complex semisimple Lie group G. Then an alcove a of the associated affine Weyl group parameterizes conjugacy classes in K. The following is an analogue of Problem Q2.

Problem 6.8. Find necessary and sufficient conditions on elements $\lambda, \mu, \nu \in a$ in order that there exist elements $A, B, C \in K$ whose projections to a are λ, μ and ν, respectively, so that $ABC = 1$.

This problem was solved independently by S. Agnihotri and C. Woodward [AW] and P. Belkale ([AW], [Be1]) in the case $K = U(n)$. This solution was generalized to all simple groups G by C. Teleman and C. Woodward in [TW]. The solution is given in a form of *nonhomogeneous* linear inequalities analogous to $(*_{\zeta;\vec{\eta}})$, where Schubert calculus is replaced with *quantum Schubert calculus*. An analogue of Horn recurrence formula for this problem was established by P. Belkale in [Be3].

Problem 6.9. 1. Solve the analogue of Problem 1.1 for compact symmetric spaces.

2. Is there an analogue of Problem Q3 in the setting of compact groups?

In the context of compact Lie groups, Problem Q4 generalizes to the problem about *product structure of the fusion ring $R_\ell(G)$ at level ℓ*, which is a certain quotient of the representation ring of the group G. For the characters $\mathrm{ch}(V_\lambda)$, $\mathrm{ch}(V_\mu)$, $\mathrm{ch}(V_\nu)$ in the ring $R_\ell(G)$ consider the decomposition of the triple product:

$$\mathrm{ch}(V_\lambda) \cdot \mathrm{ch}(V_\mu) \cdot \mathrm{ch}(V_\nu) = \sum_\delta \tilde{n}_{\lambda,\mu,\nu,\ell}(\delta)\mathrm{ch}(V_\delta).$$

Problem 6.10. Give necessary and sufficient conditions on λ, μ, ν in order that

$$\tilde{n}_{\lambda,\mu,\nu,\ell}(0) \neq 0.$$

P. Belkale proved in [Be3] an analogue of the Knutson–Tao saturation theorem for Problem 6.10, thereby establishing equivalence between Problem 6.10 and the multiplicative Problem 6.8:

Theorem 6.11. *For λ, μ, ν so that $\lambda + \mu + \nu \in Q(R)$,*

$$\tilde{n}(N\lambda, N\mu, N\nu, N\ell) \neq 0, \quad \text{for some } N \in \mathbb{N},$$

if and only if

$$\tilde{n}(\lambda, \mu, \nu, \ell) \neq 0.$$

Conjecture 6.12 (C. Woodward). The above saturation theorem holds for all simply-laced groups.

References

[AW] Agnihotri, S., and Woodward, C., Eigenvalues of products of unitary matrices and quantum Schubert calculus. *Math. Res. Lett.* **5** (1998), 817–836.

[AMW] Alekseev, A., Meinrenken, E., and Woodward, C., Linearization of Poisson actions and singular values of matrix products. *Ann. Inst. Fourier* (*Grenoble*) **51** (2001), no. 6, 1691–1717.

[Ba] Ballmann, W., *Lectures on spaces of nonpositive curvature*. With an appendix by Misha Brin, DMV Seminar 25, Birkhäuser, Basel 1995.

[B] Balser, A., Polygons with prescribed Gauss map in Hadamard spaces and Euclidean buildings. *Canad. Math. Bull.*, to appear.

[Be1] Belkale, P., Local systems on $\mathbb{P}^1$-S for S a finite set. *Compositio Math.* **129** (2001), 67–86.

[Be2] Belkale, P., Geometric proofs of Horn and Saturation conjectures. *J. Algebraic Geom.* **15** (1) (2006), 133–173.

[Be3] Belkale, P., Quantum generalization of Horn and Saturation Conjectures. Preprint math.AG/0303013, submitted.

[Be4] Belkale, P., Extremal unitary local systems on $\mathbb{P} - \{p_1, \ldots, p_s\}$. In *Proceedings of the International Colloquium on Algebraic Groups*, Tata Institute 2004, to appear.

[BK] Belkale, P., Kumar, S., Eigenvalue problem and a new product in cohomology of flag varieties. Preprint, math.AG/0407034, 2004.

[BS] Berenstein, A., and Sjamaar, R., Coadjoint orbits, moment polytopes, and the Hilbert-Mumford criterion. *J. Amer. Math. Soc.* **13** (2000), no. 2, 433–466.

[BZ] Berenstein, A., and Zelevinsky, A., Tensor product multiplicities, canonical bases and totally positive varieties. *Invent. Math.* **143** (2001), 77–128.

[BCG] Besson, G., Courtois, G., and Gallot, S., Lemme de Schwarz réel et applications géométriques. *Acta Math.* **183** (1999), 145–169.

[DW] Derksen, H., Weyman, J., Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients. *J. Amer. Math. Soc.* **13** (2000), 467–479.

[EL] Evens, S., Lu, J.-H., Thompson's conjecture for real semisimple Lie groups. In *The breadth of symplectic and Poisson geometry*, Progr. Math. 232, Birkhäuser, Boston 2005, 121–137.

[Fu] Fulton, W., Eigenvalues, invariant factors, highest weights, and Schubert calculus. *Bull. Amer. Math. Soc.* (*N.S.*) **37** (2000), no. 3, 209–249.

[G] Gauss, F., Letter to W. Bolyai. In *Collected Works*, Vol. 8, Georg Olms Verlag, Hildesheim 1973, 222–223.

[GL] Gaussent, S., and Littelmann, P., LS-Galleries, the path model and MV-cycles. *Duke Math. J.* **127** (2005), 35–88.

[Ha1] Haines, T., Structure constants for Hecke and representations rings. *Int. Math. Res. Not.* **39** (2003), 2103–2119.

[Ha2] Haines, T., Equidimensionality of convolution morphisms and applications to saturation problems. Preprint 2004, submitted.

[HKM] Haines, T., Kapovich, M., Millson, J. J., Ideal quadrilaterals in Euclidean buildings, constant term maps for spherical Hecke rings and branching to Levi subgroups. Preprint.

[KM1] Kapovich, M., Millson, J. J., A path model for geodesic in Euclidean buildings and its applications to the representation theory. Preprint, November 2004, submitted.

[KM2] Kapovich, M., Millson, J. J., Structure of the tensor product semigroup. *Asian Math. J.* (Chern memorial volume), to appear.

[KLM1] Kapovich, M., Leeb, B., Millson, J. J., Convex functions on symmetric spaces, side lengths of polygons and the stability inequalities for weighted configurations at infinity. Preprint, 2005, submitted.

[KLM2] Kapovich, M., Leeb, B., Millson, J. J., Polygons in buildings and their refined side-lengths. Preprint, 2005, submitted.

[KLM3] Kapovich, M., Leeb, B., Millson, J. J., The generalized triangle inequalities in symmetric spaces and buildings with applications to algebra. MPI Preprint, 2002; *Mem. Amer. Math. Soc.*, to appear.

[K1] Klein, T., The multiplication of Schur functions and extensions of p-modules. *J. London Math. Soc.* **43** (1968), 280–284.

[K2] Klein, T., The Hall polynomial. *J. Algebra* **12** (1969), 61–78.

[Kly1] Klyachko, A., Stable bundles, representation theory and Hermitian operators. *Selecta Math.* **4** (1998), 419–445.

[Kly2] Klyachko, A., Random walks on symmetric spaces and inequalities for matrix spectra. *Linear Algebra Appl.* **319** (Special Issue) (2000), 37–59.

[KT] Knutson, A., and Tao, T., The honeycomb model of $\mathrm{GL}_n(\mathbb{C})$ tensor products. I. Proof of the saturation conjecture. *J. Amer. Math. Soc.* **12** (1999), 1055–1090.

[KTW] Knutson, A., Tao, T., and Woodward C., The honeycomb model of $\mathrm{GL}_n(\mathbb{C})$ tensor products. II. Puzzles determine the facets of the Littlewood-Richardson cone. *J. Amer. Math. Soc.* **17** (2004), 19–48.

[KuLM] Kumar, S., Leeb, B., Millson, J. J., The generalized triangle inequalities for rank 3 symmetric space of noncompact type. In *Explorations in complex and Riemannian geometry* (Papers dedicated to Robert Greene), Contemp. Math. 332, Amer. Math. Soc., Providence, RI, 2003, 171–195.

[La] Laskowski, M. C. , An application of Kochen's Theorem. *J. Symbolic Logic* **68** (2003), no. 4, 1181–1188.

[Li] Littelmann, P., Paths and root operators in representation theory. *Ann. of Math.* (2) **142** (1995), no. 3, 499–525.

[MS1] Mulmuley, K., and Sohoni, M., Geometric complexity theory, P vs. NP and explicit obstructions. In *Advances in Algebra and Geometry* (ed. by C. Musili), Hindustan Book Agency, New Delhi 2003, 239–261.

[MS2] Mulmuley, K., and Sohoni, M., Geometric complexity theory III: on deciding positivity of Littlewood-Richardson coefficients. ArXiv preprint cs.CC/0501076.

[TW] Teleman, C., Woodward, C., Parabolic bundles, products of conjugacy classes and quantum cohomology. *Ann. Inst. Fourier* (*Grenoble*) **53** (2003), 713–748.

Department of Mathematics, 1 Shields Ave., University of California, Davis, CA 95616, U.S.A.

E-mail: kapovich@math.ucdavis.edu

The asymptotic geometry of negatively curved spaces: uniformization, geometrization and rigidity

Bruce Kleiner*

Abstract. This is a survey of recent developments at the interface between quasiconformal analysis and the asymptotic geometry of Gromov hyperbolic groups. The main theme is the extension of Mostow rigidity and related theorems to a broader class of hyperbolic groups, using recently developed analytic structure of the boundary.

Mathematics Subject Classification (2000). Primary 30C65; Secondary 20F57.

Keywords. Quasiconformal geometry, geometric group theory, rigidity.

1. Introduction

The celebrated Mostow rigidity theorem [49] states that if X and X' are symmetric spaces of noncompact type with no de Rham factors isomorphic to the hyperbolic plane, and $\Gamma \subset \mathrm{Isom}(X)$, $\Gamma' \subset \mathrm{Isom}(X')$ are uniform lattices, then any isomorphism $\Gamma \to \Gamma'$ between the lattices is the restriction of a Lie group isomorphism $\mathrm{Isom}(X) \to \mathrm{Isom}(X')$. This theorem and its proof have many important implications, and have inspired numerous generalizations and variants (e.g. [53], [47], [32], [59], [57], [27], [35], [51], [44], [2]) most of which concern lattices in semi-simple groups. Mostow's proof was based on the asymptotic geometry of symmetric spaces, more specifically the quasiconformal or combinatorial structure of the boundary at infinity. Recent developments in analysis on metric spaces and quasiconformal geometry have begun to create the technical framework needed to implement Mostow's proof in a much more general context, yielding new Mostow-type rigidity theorems. The goal of this article is to survey some of these developments and their group theoretic applications.

Organization of the paper. Section 2 discusses some general issues in geometric group theory. Section 3 covers basic facts about Gromov hyperbolic spaces and their boundaries. Section 4 reviews a selection of recent work related to quasiconformal geometry in a metric space setting, and some applications of this are covered in Section 5. Quasiconformal uniformization is discussed in Section 6, and quasiconformal geometrization in Section 7. The last section presents some open problems.

*The author gratefully acknowledges support from NSF grants DMS-0224104 and DMS-0505610.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The survey by Mario Bonk in these Proceedings provides a complementary viewpoint on some of the material presented here.

Acknowledgements. I would like to thank Mario Bonk, Marc Bourdon, Juha Heinonen, and John Mackay for numerous helpful comments on an earlier version of this article.

2. Rigidity and geometrization in geometric group theory

The ideas presented in this section are strongly influenced by the work of Gromov [32], [31], [34].

The asymptotic viewpoint in geometric group theory. One of the guiding themes in geometric group theory is that one can often understand the algebraic structure of a group by finding the right geometric realization of the group as a group of isometries. For instance it is very constructive to think of nonabelian free groups as groups acting on trees, and lattices in semi-simple Lie groups as groups of isometries acting on the associated symmetric space. Every finitely generated group has a plentiful supply of isometric actions, namely the actions on its Cayley graphs. Recall that if Σ is a finite generating set for a group G, then the *Cayley graph* of (G, Σ), denoted $\mathrm{Cayley}(G, \Sigma)$, is the graph with vertex set G, in which two group elements $g, g' \in G$ are joined by an edge if and only if $g = g'\sigma$ for some $\sigma \in \Sigma$. The action of G on itself by left translation extends to an action $G \curvearrowright \mathrm{Cayley}(G, \Sigma)$ by graph isomorphisms; one may view this as an isometric action by equipping the Cayley graph with the path metric where each edge has length 1. As a tool for understanding the original group G, Cayley graphs have a drawback: there are too many of them. Nonetheless their asymptotic, or large-scale, structures are all the same. To formalize this idea, we now recall a few definitions.

Definition 2.1. A (possibly discontinuous) map $f : X \to X'$ between metric spaces is a *quasi-isometry* if there are constants L, A such that for every $x_1, x_2 \in X$,

$$\frac{1}{L} d(x_1, x_2) - A \leq d(f(x_1), f(x_2)) \leq L d(x_1, x_2) + A, \tag{2.1}$$

and every $x' \in X'$ lies within distance at most A from a point in the image of f. Here and elsewhere we will use the generic letter "d" for metric space distance functions when it is clear in which metric space distances are being measured. Two metric spaces are *quasi-isometric* if there exists a quasi-isometry from one to another; this defines an equivalence relation on the collection of metric spaces.

A metric space is *proper* if every closed ball is compact. An isometric action $G \curvearrowright X$ on a metric space X is *discrete* if for every ball $B = B(x, r) \subset X$, the set

$$\{g \in G \mid gB \cap B \neq \emptyset\} \tag{2.2}$$

is finite, and *cocompact* if there is a compact subset $K \subset X$ such that

$$X = \bigcup_{g \in G} gK. \tag{2.3}$$

The fundamental lemma of geometric group theory ties these together:

Lemma 2.2. *Suppose G is a finitely generated group, X, X' are proper geodesic spaces, and $G \overset{\rho}{\curvearrowright} X$, $G \overset{\rho'}{\curvearrowright} X'$ are two discrete, cocompact, and isometric actions of G. Then X is quasi-isometric to X'. Moreover one may find a quasi-isometry $f: X \to X'$ which is "quasi-equivariant" in the sense that there is a constant D such that*

$$d(f(gx), gf(x)) < D \tag{2.4}$$

for all $g \in G$, $x \in X$.

If Σ is a finite generating set for a group G, then the action $G \curvearrowright \mathrm{Cayley}(G, \Sigma)$ is discrete, cocompact, and isometric; therefore the collection of actions covered by the lemma is nonempty for any finitely generated group G. This means that there is a well-defined quasi-isometry class of geodesic metric spaces associated with each finitely generated group G, which we will refer to as the quasi-isometry class of G. Another implication of the lemma is that the universal cover of a compact, connected Riemannian manifold M (equipped with the Riemannian distance function) is quasi-isometric to $\pi_1(M)$; this follows from Lemma 2.2 because the deck group action $\pi_1(M) \curvearrowright \tilde{M}$ is discrete, cocompact, and isometric. Thus in addition to Cayley graphs, one has an abundance of other metric spaces representing the quasi-isometry class of the group.

The asymptotic approach to geometric group theory is to study groups by identifying quasi-isometry invariant structure in their quasi-isometry class.

Some asymptotic problems. One of the first quasi-isometry invariants of a metric space X is its *quasi-isometry group*, denoted $\mathrm{QI}(X)$. This is defined to be the collection of equivalence classes of quasi-isometries $f: X \to X$, where two quasi-isometries f, f' are declared to be equivalent if and only if their supremum distance

$$d(f, f') := \sup_{x \in X} d(f(x), f'(x))$$

is finite, and the group law is induced by composition of quasi-isometries. The group $\mathrm{QI}(X)$ is quasi-isometry invariant because a quasi-isometry $X \to X'$ between two metric spaces induces an isomorphism $\mathrm{QI}(X) \to \mathrm{QI}(X')$ by "quasi-conjugation". By Lemma 2.2 it therefore makes sense to speak of the quasi-isometry group of a finitely generated group G.

We now discuss several problems which are central in geometric group theory.

Question 2.3.

A (QI classification of groups). What are the finitely generated groups in a given quasi-isometry class?

B (Classification of QI's). Given a finitely generated group G, what is the quasi-isometry group of G?

C (Uniformization/Recognition). Given a model group, find a criterion for deciding when another group is quasi-isometric to it.

Question A has led to some remarkable mathematics. In each case where it was resolved successfully, new geometric, analytic, combinatorial, or topological ingredients were brought into play. Here are few examples:

• A finitely generated group is quasi-isometric to $\mathbb{Z}^n$ if and only if it is virtually $\mathbb{Z}^n$, i.e. it contains a finite index subgroup isomorphic to $\mathbb{Z}^n$. The proof relies on Gromov's theorem on groups of polynomial growth [33], Pansu's description of asymptotic cones of nilpotent groups [50], and Bass's formula for the growth of a nilpotent group [1]. We note that a new proof was found recently by Shalom [55]; it uses a simpler approach, but it is still far from elementary.

• A finitely generated group quasi-isometric to a free group F_k is virtually free. This is due to Gromov, and uses Stallings' theorem on ends of groups [56].

• A finitely generated group quasi-isometric to a symmetric space X of noncompact type admits a discrete, cocompact, isometric action on X. This follows from combined work of Sullivan, Gromov, Tukia, Pansu, Kleiner–Leeb, and Gabai, Casson–Jungreis. The proofs involve asymptotic geometry – quasiconformal structure on boundaries, asymptotic cones, and Tits geometry.

Question A is related to Question B because if a group G' is quasi-isometric to G, then there is a homomorphism $G' \to \mathrm{QI}(G') \simeq \mathrm{QI}(G)$, and one can approach Question A by studying subgroups of $\mathrm{QI}(G)$.

In most of the cases when Question B has been answered satisfactorily, the approach is to identify a "canonical" or "optimal" metric space X which is quasi-isometric to G, and then argue that the homomorphism $\mathrm{Isom}(X) \to \mathrm{QI}(X)$ is an isomorphism, i.e. that every quasi-isometry of X lies at finite distance from a unique isometry [51], [44], [12]. When this happens, one says that the metric space X is *quasi-isometrically rigid*. This leads to:

Question 2.4. When can one find a quasi-isometrically rigid space in the quasi-isometry class of a given finitely generated group G?

In general, an answer to this question has two parts. The first part is geometrization: finding a candidate for the optimal/rigid metric space in the given quasi-isometry class. In most of the earlier rigidity theorems, geometrization was easy because there was an obvious model space to consider. The second part is to show that the candidate space is actually rigid, which requires one to exploit appropriate asymptotic structure.

Quasi-isometric rigidity may fail because the quasi-isometry group is too large, which is what happens for $\mathbb{Z}^n$, free groups F_k, or lattices in $\mathrm{Isom}(\mathbb{H}^n)$. Nonetheless in

many cases a weaker form of rigidity survives: there is a proper metric space X in the quasi-isometry class such that any group in this class admits – possibly after passing to a finite index subgroup – a discrete, cocompact, isometric action on X. To put it another way: any group quasi-isometric to X is virtually a lattice in $\mathrm{Isom}(X)$, when one topologizes $\mathrm{Isom}(X)$ using the compact-open topology. So although individual quasi-isometries are not rigid, sufficiently large groups of quasi-isometries may turn out to be rigid. When this weaker form of rigidity holds, one may ask if the analog of Mostow rigidity is true.

Question C has a satisfactory answer for only a few classes of groups, such as free groups, free abelian groups, and surface groups. For instance, a finitely generated group G is virtually free if for every Cayley graph $\mathcal{G}$ there are constants r, R such that any two points $x, y \in \mathcal{G}$ at distance at least R lie in distinct components of the complement of some r-ball $B(z, r) \subset \mathcal{G}$. An important unresolved case of Question C, which is tied to conjectures in 3-manifold topology, is to find a characterization of groups quasi-isometric to hyperbolic 3-space $\mathbb{H}^3$. This is discussed in Section 6.

In the remainder of this paper, we will focus on these questions and related mathematics in the context of negatively curved manifolds, or more generally Gromov hyperbolic spaces.

3. Gromov hyperbolic spaces and their boundaries

In this section we review some facts about Gromov hyperbolicity, see [31], [30], [16], [34], [41]. Gromov hyperbolic spaces form a robust class of metric spaces to which much of the theory of negatively curved Riemannian manifolds applies. They have a boundary at infinity, which plays an essential role in rigidity applications. Much of their asymptotic structure is encoded in the quasiconformal (or quasi-Möbius) structure of the boundary; this fact enables one to exploit the analytic theory of quasiconformal homeomorphisms.

We recall that a geodesic metric space X is *Gromov hyperbolic* if there is a constant δ such that every geodesic triangle in X is δ-*thin*, in other words, each side lies within the δ-neighborhood of the union of the other two sides. Gromov hyperbolicity is a quasi-isometry invariant property for geodesic metric spaces [16]. A finitely generated group is *Gromov hyperbolic* if its quasi-isometry class is Gromov hyperbolic, see Lemma 2.2 and the ensuing commentary.

Except when it is explicitly stated to the contrary, for the remainder of this paper X will denote a proper Gromov hyperbolic geodesic metric space and δ its hyperbolicity constant.

Prime examples of Gromov hyperbolic spaces are complete simply-connected Riemannian manifolds of sectional curvature bounded above by a constant $\kappa < 0$, equipped with their Riemannian distance functions, and more generally, simply-connected Alexandrov spaces of curvature $\leq \kappa$ [16]. Metric trees and piecewise

Euclidean polyhedra satisfying appropriate link conditions also provide many examples of group theoretic interest [31], [30], [16].

Two geodesic rays $\gamma_1, \gamma_2 \colon [0, \infty) \to X$ are *asymptotic* if the Hausdorff distance between their images is finite; this defines an equivalence relation on the collection of geodesic rays in X. The *boundary* of X, denoted ∂X, is the collection of equivalence classes. Pick $p \in X$. Given $[\gamma] \in \partial X$, there is a unit speed geodesic ray starting from p which is asymptotic to γ; thus one may identify ∂X with the collection of asymptote classes of geodesics rays starting at p. Given unit speed geodesic rays $\gamma_1, \gamma_2 \colon [0, \infty) \to X$ starting from p, their *Gromov overlap* is defined to be

$$\langle \gamma_1 \mid \gamma_2 \rangle := \lim_{t \to \infty} \frac{1}{2} \left(2t - d(\gamma_1(t), \gamma_2(t))\right) \in [0, \infty]. \tag{3.1}$$

To within an additive error comparable to the hyperbolicity constant δ, the overlap of two rays is the infimal $t \in \mathbb{R}$ such that the distance from $\gamma_1(t)$ to $\gamma_2(t)$ is $> \delta$.

Visual metrics. A *visual metric* on ∂X is a metric ρ such that for some constants $a > 0$ and C (a is called the visual parameter of ρ), and some $p \in X$,

$$\frac{1}{C} e^{-a\langle \gamma_1 | \gamma_2 \rangle} \leq \rho(\gamma_1, \gamma_2) \leq C e^{-a\langle \gamma_1 | \gamma_2 \rangle} \tag{3.2}$$

for every pair of rays γ_1 and γ_2 starting at p. This condition is independent of the choice of basepoint used to define the overlap. When a is small compared to $1/\delta$, visual metrics with visual parameter a always exist.

Henceforth when we refer to the boundary ∂X, we will mean the set ∂X equipped with some visual metric, unless otherwise stated. Here are some examples of visual metrics (note that the assertions apply to a particular choice of visual metric):

• The boundary of $\mathbb{H}^n$ is bi-Lipschitz homeomorphic to the sphere S^{n-1} equipped with the usual metric.

• The boundary of complex hyperbolic space $\mathbb{CH}^n$ is bi-Lipschitz homeomorphic to S^{2n-1} equipped with the usual Carnot metric. We recall that this metric is defined as follows. Let Δ be the usual contact structure on S^{2n-1} induced by the embedding of S^{2n-1} into $\mathbb{C}^n$. The Carnot distance between two points $p, q \in S^{2n-1}$ is the infimum of the lengths of integral curves of Δ which join p to q. This metric on the $(2n-1)$-sphere has Hausdorff dimension $2n$.

• Let T be a trivalent simplicial tree where each edge has length 1. Then ∂T is a Cantor set.

• Let M be a compact Riemannian 3-manifold of constant sectional curvature -1 with nonempty totally geodesic boundary, and let X be its universal cover equipped with the Riemannian distance function. Then X isometrically embeds in $\mathbb{H}^3$ as a convex subset C bounded by a countable collection of disjoint totally geodesic planes, and ∂X may be identified with the limit set of C in the sphere at infinity of $\mathbb{H}^3$. The limit set is obtained by removing a countable disjoint collection of round spherical caps from the 2-sphere, and is homeomorphic to the Sierpiński carpet.

There is not much one can say about boundaries of Gromov hyperbolic spaces in general since every compact doubling metric space is isometric to the boundary of a proper Gromov hyperbolic space equipped with a visual metric, see [13, Section 2]. However, if X admits a discrete, cocompact, isometric action, then the boundary structure is much more restricted – it is *approximately self-similar* in the following sense. There are constants $L,\ D > 0$ such that if $B(p,r) \subset \partial X$ is a ball with $0 < r \leq \operatorname{diam}(\partial X)$, then there is an open subset $U \subset \partial X$ of diameter $> D$ which is L-bi-Lipschitz homeomorphic to the rescaled ball $\frac{1}{r}B(p,r)$. This approximate self-similarity has many implications for the topology and geometry of ∂X. Before stating them, we need several definitions.

Definition 3.1. A metric space Z is *Ahlfors Q-regular* if there is a constant C such that the Q-dimensional Hausdorff measure of any r-ball B satisfies

$$\frac{1}{C}r^Q \leq \mathcal{H}^Q(B) \leq Cr^Q, \tag{3.3}$$

provided $r \leq \operatorname{diam}(Z)$.

Note that an Ahlfors Q-regular space has Hausdorff dimension Q. Examples of Ahlfors regular spaces are $\mathbb{R}^n$ and S^{2n-1} equipped with standard Carnot metric.

The next definitions are scale-invariant, quantitative versions of standard topological conditions.

Definition 3.2. A metric space Z is *linearly locally contractible* if there is a constant $\lambda > 0$ such that for all $x \in Z$, $0 < r \leq \operatorname{diam}(Z)$, the inclusion

$$B(x, \lambda r) \to B(x, r) \tag{3.4}$$

is null-homotopic. Linear local contractibility excludes examples where the metric topology looks worse and worse at smaller and smaller scales. For instance, a metric on a two sphere which has a sequence of "fingers" of smaller and smaller diameter, for which the ratio of the diameter to the circumference tends to infinity, will not be linearly locally contractible.

A metric space Z is *uniformly perfect* if there is a constant $\lambda > 0$ such that if $p \in Z$, $0 < r \leq \operatorname{diam}(Z)$, then $B(p,r) \setminus B(p, \lambda r)$ is nonempty.

A metric space Z is *linearly locally connected*, or LLC (not to be confused with "linearly locally contractible"), if there is an L such that for all $p \in Z$, $0 < r \leq \operatorname{diam}(X)$, the inclusions

$$B(p,r) \to B(p, Lr), \quad X \setminus B(p,r) \longrightarrow X \setminus B\left(p, \frac{r}{L}\right) \tag{3.5}$$

induce the zero homomorphism on reduced 0-dimensional homology. This is a standard type of condition in quasiconformal geometry, and is a quantitative version of local connectedness and absence of local cut points.

Theorem 3.3 (Structure of the boundary). *If X is a proper, geodesic, Gromov hyperbolic space which admits a discrete, cocompact, isometric action, then:*

1. *∂X is either empty, has two elements, or is uniformly perfect.*
2. *∂X is Ahlfors Q-regular for some Q* ([26]).
3. *If ∂X is connected, it is LLC* ([4], [15], [58], [7]).
4. *If ∂X is homeomorphic to S^k, then ∂X is linearly locally contractible.*
5. *If X is a contractible n-manifold, then ∂X is a homology $(n-1)$-manifold with the homology of the $(n-1)$-sphere* [4], [3]. *When $n \leq 3$ then ∂X is a sphere.*

Induced homeomorphisms between boundaries. In applications to group theory and rigidity, a key property of the boundary is that quasi-isometries between Gromov hyperbolic spaces induce homeomorphisms between their boundaries. To formulate this more precisely we need the following definition, which is due to Väisälä.

Definition 3.4 ([63]). The *cross-ratio*, $[z_1, z_2, z_3, z_4]$, of a 4-tuple of distinct points (z_1, z_2, z_3, z_4) in a metric space Z is the quantity

$$[z_1, z_2, z_3, z_4] := \frac{d(z_1, z_3)d(z_2, z_4)}{d(z_1, z_4)d(z_2, z_3)}. \tag{3.6}$$

This is a metric space version of the familiar cross-ratio in complex analysis. A homeomorphism $\phi\colon Z \to Z'$ between metric spaces is *quasi-Möbius* if there is a homeomorphism $\eta\colon [0,\infty) \to [0,\infty)$ such that for every quadruple of distinct points $(z_1, z_2, z_3, z_4) \in Z$,

$$[\phi(z_1), \phi(z_2), \phi(z_3), \phi(z_4)] \leq \eta([z_1, z_2, z_3, z_4]). \tag{3.7}$$

Such a homeomorphism η is called a *distortion function* for the homeomorphism ϕ. Intuitively, a homeomorphism is quasi-Möbius if it distorts cross-ratios in a controlled way.

Compositions and inverses of quasi-Möbius homeomorphisms are quasi-Möbius, so every metric space Z has an associated group of quasi-Möbius homeomorphisms, which we denote by $\mathrm{QM}(Z)$.

Theorem 3.5 ([52]). *Every quasi-isometry $f\colon X \to X'$ between Gromov hyperbolic spaces induces a quasi-Möbius homeomorphism $\partial f\colon \partial X \to \partial X'$ between their boundaries. The distortion function η for ∂f can be chosen to depend only on the hyperbolicity constants of X and X', the constants for the visual metrics on ∂X and $\partial X'$, and the quasi-isometry constants of f.*

In particular, the full isometry group of a Gromov hyperbolic space acts on its boundary as a group of quasi-Möbius homeomorphisms with uniform distortion function.

The proof of Theorem 3.5 has two ingredients. The first is the "Morse Lemma", which implies that f maps each geodesic segment (respectively ray) in X to within controlled Hausdorff distance of a geodesic segment (respectively ray) in X'. This yields the induced set-theoretic bijection $\partial f\colon \partial X \to \partial X'$. To verify that ∂f is quasi-Möbius, the idea is to associate to each 4-tuple of points in ∂X a configuration in X consisting of a pair of geodesics and a shortest geodesic segment connecting them. The cross-ratio is determined by the length of the connecting geodesic segment, and the latter is preserved to within a factor by the quasi-isometry.

Under mild assumptions, for instance if X is quasi-isometric to a nonelementary hyperbolic group (a hyperbolic group which does not contain a finite index cyclic subgroup), the homomorphism $\mathrm{QI}(X) \to \mathrm{QM}(\partial X)$ given by taking boundary homeomorphisms is an isomorphism. Thus one can translate questions about quasi-isometries into questions about quasi-Möbius homeomorphisms of the boundary. This has the additional advantage of eliminating some ambiguity (an equivalence class of quasi-isometries is replaced by a single quasi-Möbius homeomorphism), as well as providing extra structure – the group $\mathrm{QM}(Z)$ has a natural topology.

Questions A–C in Section 2 translate to:

A′. What are the hyperbolic groups whose boundary is quasi-Möbius homeomorphic to a given metric space Z?

B′ What is the group of quasi-Möbius homeomorphisms of the metric space ∂G?

C′. Given a boundary ∂G, how can one tell if another boundary (or space) is quasi-Möbius homeomorphic to it?

The latter two questions make perfect sense and are interesting for spaces other than boundaries, e.g. self-similar spaces like the standard square Sierpinski carpet or the Menger sponge.

4. Quasiconformal homeomorphisms

This section presents some recent results on quasiconformal homeomorphisms. The material was selected for its applicability to rigidity and group theoretic problems, and does not represent a balanced overview. See [38], [37], [61], [25] for more discussion. The somewhat separate topic of uniformization is discussed in Section 6.

We begin with some definitions.

Definition 4.1. Let $f\colon X \to X'$ be a homeomorphism between metric spaces, and $p \in X$. The *dilatation of f at p* is

$$H(f, p) := \limsup_{r\to 0} \frac{\sup\{d(f(x), f(p)) \mid x \in B(p, r)\}}{\inf\{d(f(x), f(p)) \mid x \notin B(p, r)\}}. \tag{4.1}$$

The homeomorphism f is *C-quasiconformal* if $H(f, p) \le C$ for every $p \in X$, and *quasiconformal* if it is C-quasiconformal for some C.

Heuristically, a homeomorphism is quasiconformal if it maps infinitesimal balls to sets of controlled eccentricity.

Definition 4.2. Let (Z, μ) be a metric space equipped with a Borel measure μ, and $p \geq 1$. If Γ is a collection of paths in Z (i.e. continuous maps $[0, 1] \to Z$), then a Borel measurable function $\rho\colon Z \to [0, \infty]$ is Γ-*admissible* if

$$\int_\gamma \rho \, ds \geq 1 \tag{4.2}$$

for every rectifiable path $\gamma \in \Gamma$, where ds denotes the arclength measure. The *p-modulus of* Γ is the infimum of the quantities

$$\int_Z \rho^p d\mu \tag{4.3}$$

where ρ ranges over all Γ-admissible functions on Z. If E, F are subsets of Z, then the *p-modulus of* (E, F), denoted $\mathrm{Mod}_p(E, F)$, is the p-modulus of the collection of paths running from E to F. When Z is an Ahlfors Q-regular space and no measure is specified, we will be using Q-dimensional Hausdorff measure by default.

Modulus is an old and important tool in conformal and quasiconformal geometry, due to its conformal invariance and the fact that it permits one to relate infinitesimal with global behavior of homeomorphisms. One checks by a change of variable computation that a conformal diffeomorphism $M \to M'$ between Riemannian manifolds preserves modulus of curve families, and that a diffeomorphism which preserves modulus is conformal. More generally, the effect of a diffeomorphism on modulus can be controlled by its dilatation.

Quasiconformal homeomorphisms between domains in $\mathbb{R}^n$ have a number of important regularity properties.

Theorem 4.3 ([62]). *Let $f\colon U \to U'$ be a quasiconformal homeomorphism between domains in $\mathbb{R}^n$, $n \geq 2$. Then f belongs to the Sobolev space $W^{1,n}_{\mathrm{loc}}(U, U')$, it is differentiable almost everywhere, and maps sets of measure zero to sets of measure zero. Furthermore, f is ACL; this means that for every direction $v \in \mathbb{R}^n$, for almost every line L parallel to v, the restriction of f to $U \cap L$ is absolutely continuous with respect to 1-dimensional Hausdorff measure $\mathcal{H}^1$. Furthermore, inverses and compositions of quasiconformal homeomorphisms between such domains are quasiconformal.*

The following theorem shows that there are many characterizations of quasiconformal homeomorphisms. This is important because different characterizations are useful in different situations, and also because it demonstrates that this notion is robust.

Theorem 4.4 (see [62]). *Let M and M' be Riemannian manifolds of dimension $n \geq 2$. Then the following conditions on a homeomorphism $f\colon M \to M'$ are quantitatively equivalent:*

1. *f is C-quasiconformal.*
2. *f distorts n-modulus of curve families by a factor at most L.*
3. *f belongs to the Sobolev space $W^{1,n}_{\mathrm{loc}}(M, M')$ and the distributional derivative Df satisfies $|Df|^n \leq C\,\mathrm{Jac}(f)$ almost everywhere, for some C, where $\mathrm{Jac}(f)$ denotes the Jacobian of f with respect to the Riemannian structures on M and M' (the infinitesimal volume distortion factor).*

If one assumes in addition that M and M' are compact, then one may add quasi-Möbius to this list of quantitatively equivalent conditions:

4. *f is η-quasi-Möbius.*

These two theorems show that quasiconformal homeomorphisms are very nicely behaved in a Riemannian context, when $n \geq 2$. Unfortunately, for general metric spaces, or even general Ahlfors regular metric spaces, Definition 4.1 is not necessarily equivalent to the quasi-Möbius condition, and does not lead to a useful theory. Examples show that it is necessary to impose further conditions on the structure of the metric space in order to prove anything akin to Theorem 4.4. After earlier work on Carnot–Carathéodory geometry [49], [45], [48], the breakthrough paper [38] introduced natural conditions for precisely this purpose, and by now they have been shown to imply most of the properties above. The essential requirement on the space is a quantitative link between infinitesimal structure and global structure, which is expressed by a relation between moduli of pairs (E, F) and their relative distance. The *relative distance* between two subsets E, F of a metric space is

$$\Delta(E, F) := \frac{\mathrm{dist}(E, F)}{\min\left(\mathrm{diam}(E), \mathrm{diam}(F)\right)}, \tag{4.4}$$

where

$$\mathrm{dist}(E, F) := \inf\{d(x, y) \mid x \in E,\ y \in F\}. \tag{4.5}$$

This is a scale invariant measure of the separation between two subsets.

Definition 4.5. Let Z be an Ahlfors Q-regular metric space. Then Z is *Q-Loewner* if there is a positive decreasing function $\phi\colon (0, \infty) \to (0, \infty)$ such that

$$\mathrm{Mod}_Q(E, F) \geq \phi(\Delta(E, F)), \tag{4.6}$$

whenever $E, F \subset Z$ are disjoint nondegenerate continua. We recall that a *continuum* is a compact, connected, subset, and a nondegenerate continuum is one of positive diameter. Recall that for Q-regular spaces, Q-dimensional Hausdorff measure is the default measure used to calculate modulus.

Examples of Loewner spaces include Euclidean spaces, compact connected Riemannian manifolds, Carnot groups, and complete connected Riemannian manifolds of nonnegative Ricci curvature.

We remark that for Ahlfors Q-regular spaces, one always has an upper bound on modulus of the following form. There is a function

$$\psi : [0, \infty) \to [0, \infty] \tag{4.7}$$

with $\lim_{t\to\infty} \psi(t) = 0$ such

$$\mathrm{Mod}_Q(E, F) \leq \psi(\Delta(E, F)) \tag{4.8}$$

for any pair of closed subsets E, F. Combining (4.6) and (4.8), one can say that for a Q-Loewner space, the modulus for a pair of disjoint continua is quantitatively controlled by their relative distance.

The paper [38] introduced Definition 4.5, and together with subsequent work [61], [39], [25], most of the facts in Theorems 4.3 and 4.4 have now been proved for quasiconformal homeomorphisms between Loewner spaces.

Theorem 4.6. *Let Z, Z' be compact Q-Loewner spaces for some $Q > 1$.*

1. *Quasiconformal homeomorphisms $Z \to Z'$ are absolutely continuous, i.e. map sets of measure zero to sets of measure zero.*
2. *Every quasiconformal homeomorphism $f : Z \to Z'$ is ACL in the following sense. There is a collection of paths Γ of Q-modulus zero such that if $\gamma : [0, 1] \to Z$ is a path, $\gamma \notin \Gamma$, then $f \circ \gamma : [0, 1] \to Z'$ is absolutely continuous with respect to 1-dimensional Hausdorff measure on the target.*
3. *Every quasiconformal homeomorphism $Z \to Z'$ belongs to the Sobolev space $W^{1,Q}(Z, Z')$.*
4. *For a homeomorphism $f : Z \to Z'$, conditions 1, 2, and 4 of Theorem 4.4 are quantitatively equivalent.*
5. *The inverse of a quasiconformal homeomorphism is quasiconformal.*

In the paper [61] it was shown that quasi-Möbius homeomorphisms between compact Q-regular spaces distort Q-modulus in a controlled way; in particular the Q-Loewner property is invariant under quasi-Möbius homeomorphisms.

The paper [38] also showed that when $Q > 1$, for compact Ahlfors Q-regular spaces, the Q-Loewner condition is equivalent to a $(1, Q)$-Poincaré inequality. This is an analytic condition which relates upper gradients to mean oscillation. We refer the reader to [37] for more on this topic.

Another breakthrough was the paper [25] which established a notion of differentiability for Lipschitz functions on doubling metric measure spaces satisfying a Poincaré inequality, in particular for compact Q-Loewner spaces. This has the remarkable implication that there is a *cotangent bundle* T^*Z for such metric measure spaces, which is a measurable vector bundle equipped with a canonical measurably varying fiberwise norm. Any bi-Lipschitz homeomorphism $(Z, \mu) \to (Z', \mu')$ preserving measure classes induces a derivative mapping $T^*Z' \to T^*Z$, which is a

measurable bundle isomorphism which distorts the norms by a factor controlled by the bi-Lipschitz constant. The paper [39] extended this assertion to Sobolev functions, and showed that quasiconformal homeomorphisms $Z \to Z'$ between Q-regular, Q-Loewner spaces also have a well-defined Cheeger derivative, when $Q > 1$. Using this the author showed [43] that under the same assumptions, the dilatation of a quasiconformal homeomorphism is controlled quantitatively by the dilatation of its derivative mapping $T^*Z' \to T^*Z$; the latter is defined relative to the canonical fiberwise norms on T^*Z and T^*Z'.

5. Applications to rigidity

We now discuss applications of the analytic results in the previous section to rigidity theorems.

The first is Mostow rigidity in the negatively curved case:

Theorem 5.1. *Suppose X and X' are rank* 1 *symmetric spaces of noncompact type other than $\mathbb{H}^2$, and G and G' are cocompact lattices in* $\mathrm{Isom}(X)$ *and* $\mathrm{Isom}(X)$ *respectively. Then any isomorphism $G \to G'$ extends to a Lie group isomorphism* $\mathrm{Isom}(X) \to \mathrm{Isom}(X')$.

The outline of the proof goes as follows. Identifying the two groups G and G' using the isomorphism, one obtains discrete, cocompact, isometric actions $G \curvearrowright X$ and $G \curvearrowright X'$. These induce uniformly quasi-Möbius (in fact Möbius in the $\mathbb{H}^n$ case) boundary actions $G \curvearrowright \partial X$, $G \curvearrowright \partial X'$. By Lemma 2.2 one gets a "quasi-equivariant" quasi-isometry

$$f : X \to X', \tag{5.1}$$

which has a quasi-Möbius boundary homeomorphism $\partial f : \partial X \to \partial X'$. The quasi-equivariance of f implies that ∂f is equivariant with respect to the actions $G \curvearrowright \partial X$ and $G \curvearrowright \partial X'$. By using the equivariance and the dynamics of the group action, one then argues that the derivative of ∂f, which is defined almost everywhere because ∂X and $\partial X'$ are Carnot spaces [49], [45], [48], [51], must actually be conformal almost everywhere. This implies that $\partial f = \partial h$, for a unique isometry $h : X \to X'$. It follows readily that h is G-equivariant, and induces the desired isomorphism $\mathrm{Isom}(X) \to \mathrm{Isom}(X')$.

Pansu used a similar outline to show that for rank 1 symmetric spaces other than the hyperbolic and complex hyperbolic spaces an even stronger rigidity result holds:

Theorem 5.2 ([51]). *Suppose X is a quaternionic hyperbolic space or the Cayley hyperbolic plane, and X' is a rank* 1 *symmetric space of noncompact type. Then any quasi-isometry $X \to X'$ is at bounded distance from a unique isometry.*

The proof of this theorem also uses boundary geometry. The boundary homeomorphism $\partial f : X \to X'$ for a quasi-isometry is quasiconformal, and Pansu shows

that for the spaces in question, the derivative is forced to be conformal even without invoking an equivariance assumption as in Mostow's proof.

This kind of argument was used in a "non-classical" setting in work of Bourdon, Bourdon–Pajot, and Xie, proving a Pansu-type rigidity result for Fuchsian buildings:

Theorem 5.3 ([8], [11], [12], [9], [64]). *Every quasi-isometry between Fuchsian buildings is at finite distance from an isomorphism.*

Here a *Fuchsian building X* is a special kind of 2-dimensional polyhedral complex. It is a union of subcomplexes called *apartments*, each of which is isomorphic to the Coxeter complex associated with a fixed Coxeter group acting on $\mathbb{H}^2$. We refer the reader to the papers above for the precise definition. The proof of this rigidity result also uses the quasiconformal structure on the boundary, but in this case the boundary is a Loewner space homeomorphic to the Menger sponge, and much of the theory in Section 4 is brought into play.

Another result in a spirit similar to Mostow rigidity is:

Theorem 5.4 ([57], [32], [59], [51], [22]). *Suppose G is a finitely generated group quasi-isometric to a rank* 1 *symmetric space X other than* $\mathbb{H}^2$. *Then G admits a discrete, cocompact, isometric action on X.*

The outline of the proof goes as follows. The group G acts by isometries on a Cayley graph $\mathrm{Cayley}(G, \Sigma)$, and a quasi-isometry $\mathrm{Cayley}(G, \Sigma) \to X$ allows one to "conjugate" this isometric action to a "quasi-action" by quasi-isometries $G \curvearrowright X$. Passing to the boundary, one obtains an action $G \curvearrowright \partial X$ by uniformly quasi-Möbius homeomorphisms, in particular uniformly quasiconformal homeomorphisms. By a lemma of Sullivan, this action $G \curvearrowright \partial X$ is actually conformal with respect to some bounded measurable Riemannian structure on $T^*\partial X$; recall that ∂X is a Loewner space and therefore has a Cheeger cotangent bundle $T^*\partial X$ whose fiberwise norm can be used to express the boundedness condition on the Riemannian structure. By using the dynamics of the action and a rescaling argument, one shows that modulo quasi-Möbius conjugation, this Riemannian metric can be taken to be standard. This means that the action is actually conformal in the usual sense, and is therefore induced by an isometric action $G \curvearrowright X$.

Now suppose Z and Z' are compact Q-Loewner metric spaces, where $Q > 1$. The differentiation theory for quasiconformal homeomorphisms enables one to make the following definition:

Definition 5.5. Suppose $\langle \cdot, \cdot \rangle$, $\langle \cdot, \cdot \rangle'$ are measurable Riemannian structures on T^*Z and T^*Z'. A homeomorphism $f\colon Z \to Z'$ is *conformal* with respect to these structures if it is quasiconformal and its derivative

$$Df(z)\colon (T^*_{f(z)}Z', \langle \cdot, \cdot \rangle') \to (T^*_z Z, \langle \cdot, \cdot \rangle) \tag{5.2}$$

is conformal for almost every $z \in Z$. The *conformal group of* $(Z, \langle \cdot, \cdot \rangle)$, denoted $\mathrm{Conf}(Z, \langle \cdot, \cdot \rangle)$, is the group of conformal homeomorphisms

$$(Z, \langle \cdot, \cdot \rangle) \to (Z, \langle \cdot, \cdot \rangle). \tag{5.3}$$

By the lemma of Sullivan quoted above, any countable group of uniformly quasiconformal homeomorphisms of Z is conformal with respect to some bounded measurable Riemannian structure on T^*Z. In particular, if G is a Gromov hyperbolic group whose boundary is quasi-Möbius homeomorphic to a Q-Loewner space for $Q > 1$, then G may be viewed as a group of conformal homeomorphisms in this sense. For such a group, the full conformal group $\mathrm{Conf}((Z, \langle \cdot, \cdot \rangle)$ provides a natural substitute for the ambient Lie group that one has in the case of lattices in rank 1 Lie groups. The following result shows that in this case the homomorphism $G \to \mathrm{Conf}((Z, \langle \cdot, \cdot \rangle)$ is canonically attached to G:

Theorem 5.6 (Mostow rigidity for Loewner groups [43]). *Suppose G is a Gromov hyperbolic group, and*

$$G \overset{\rho}{\curvearrowright} (Z, \langle \cdot, \cdot \rangle), \quad G \overset{\rho'}{\curvearrowright} (Z', \langle \cdot, \cdot \rangle') \tag{5.4}$$

are conformal actions of G on Loewner spaces which are topologically conjugate to the action of G on its boundary ∂G. Then ρ is conformally equivalent to ρ'.

The proof is identical to that of Theorem 5.4 until the last step, which requires one to exploit delicate infinitesimal structure of the Loewner space.

6. Uniformization

The uniformization problem for spheres. A Riemannian metric g is *bounded* if there is a C such that

$$\frac{1}{C} g_0(v, v) \leq g(v, v) \leq C g_0(v, v) \quad \text{for all } v \in TS^2. \tag{6.1}$$

We recall the following extension of the Koebe uniformization theorem to measurable conformal structures:

Theorem 6.1 (The measurable Riemann mapping theorem). *If g is a bounded measurable Riemannian metric on the 2-sphere, then g is conformally equivalent to the standard metric g_0, i.e. there is a quasiconformal homeomorphism $f : S^2 \to S^2$ such that the derivative*

$$Df(x) : (T_x S^2, g) \to (T_x S^2, g_0) \tag{6.2}$$

is conformal almost everywhere. Moreover the uniformizing homeomorphism is unique up to post-composition with Möbius transformation.

Theorem 6.1 and a version for parametrized families of Riemannian metrics are fundamental tools in Kleinian groups and complex dynamics.

It is very tempting to extend Theorem 6.1 to a more general setting. An approach based on a type of coverings ("shinglings") was introduced by Cannon [17], and further developed in [19], [20], [21], [18]. In a metric space setting, one is naturally led to the following quasi-Möbius uniformization problem:

Question 6.2. When is a metric n-sphere quasi-Möbius homeomorphic to the standard n-sphere $\mathbb{S}^n$?

Here a *metric n-sphere* means a metric space homeomorphic to the n-sphere.

One arrives at the $n = 2$ case of this question starting with Question C from Section 2, in the $\mathbb{H}^3$ case, since a geodesic space quasi-isometric to $\mathbb{H}^3$ is Gromov hyperbolic and has boundary quasi-Möbius homeomorphic to S^2. The question is also tied to one approach to:

Conjecture 6.3 (Thurston's Hyperbolization Conjecture). Every closed, aspherical, irreducible, atoroidal 3-manifold M admits a Riemannian metric of constant curvature -1.

The relation with Question 6.2 is as follows. Gabai–Meyerhoff–Thurston [29] reduced Conjecture 6.3 to showing that $\pi_1(M)$ is isomorphic to the fundamental group of a hyperbolic manifold (a closed Riemannian manifold of constant curvature -1). When $\pi_1(M)$ is Gromov hyperbolic, [4] implies that the boundary of $\pi_1(M)$ is homeomorphic to the 2-sphere. Therefore the Gromov hyperbolic case of Conjecture 6.3 is implied by:

Conjecture 6.4 (Cannon). If G is a Gromov hyperbolic group and ∂G is homeomorphic to the 2-sphere S^2, then G admits a discrete, cocompact, isometric action on hyperbolic 3-space $\mathbb{H}^3$.

By Theorem 5.4 and the converse of Theorem 3.5, it follows readily that Cannon's conjecture is equivalent to the following case of Question 6.2:

Conjecture 6.5. If G is a Gromov hyperbolic group and ∂G is homeomorphic to S^2, then it is quasi-Möbius homeomorphic to the standard 2-sphere.

Although Thurston's conjecture appears to have been solved by Perelman, Conjecture 6.4 remains very interesting – it is logically independent of the Hyperbolization Conjecture, and moreover it provides an approach to an old unsolved problem due to Wall: Is every 3-dimensional Poincaré duality group is a 3-manifold group?

Necessary conditions for uniformization. The uniformization problem above was discussed in [60], where two necessary conditions were identified. A metric n-sphere which is quasi-Möbius homeomorphic to the standard n-sphere must be doubling and linearly locally contractible. Recall that a metric space is *doubling* if there is a constant N such that every ball can be covered by at most N balls of half the radius. We note that if the boundary of a hyperbolic group is homeomorphic to a sphere, then it satisfies both of these conditions since Ahlfors regular spaces are always doubling, see Section 3. When $n = 1$ these two necessary conditions are sufficient:

Theorem 6.6 ([60]). *A doubling, linearly locally contractible metric circle is quasi-Möbius homeomorphic to the standard circle.*

However, when $n \geq 2$, the conditions are not sufficient. One can show that $\mathbb{R}^2$ with the homogeneous distance function

$$d((x_1, y_1), (x_2, y_2)) := |x_1 - x_2| + |y_1 - y_2|^{\frac{1}{2}} \tag{6.3}$$

is not locally quasi-Möbius homeomorphic to $\mathbb{R}^2$, and it is possible to construct a doubling linearly locally contractible metric on S^2 which is locally isometric to the metric (6.3) near some point.

Sufficient conditions for uniformization. Motivated by considerations relating analytic properties of a space and the existence of good parametrizations, Semmes made the following conjecture:

Conjecture 6.7 ([40]). If Z is an Ahlfors 2-regular, linearly locally contractible 2-sphere, then Z is quasi-Möbius homeomorphic to the standard 2-sphere.

This conjecture was proven in [5]. Recall that the Hausdorff dimension of any metric space is always greater than or equal to its topological dimension. For linearly locally contractible 2-spheres, one can strengthen this to the quantitative assertion that every r-ball has 2-dimensional Hausdorff measure at least comparable to r^2, for $r \leq \operatorname{diam}(Z)$. Thus the Ahlfors 2-regularity condition in the hypothesis of Conjecture 6.7 provides a competing upper bound on the Hausdorff measure; in some sense this tension is the key to the proof. We remark that this result is optimal in several respects. First, Semmes had shown in [54] that the analogous assertion is false in higher dimensions: for every $n \geq 3$ there are Ahlfors n-regular, linearly locally contractible metric n-spheres which are not quasi-Möbius homeomorphic to the standard n-sphere. Also, the conclusion cannot be strengthened to bi-Lipschitz homeomorphism, due to examples of Laakso [46]. Finally, the Ahlfors 2-regularity condition cannot be relaxed to Q-regularity because by using metrics as in (6.3) one can get examples which are Ahlfors 3-regular.

Nonetheless, one can relax the 2-regularity condition if one imposes a Loewner condition:

Theorem 6.8 ([5]). *If Z is a Q-Loewner metric 2-sphere, then $Q = 2$ and Z is quasi-Möbius homeomorphic to the standard 2-sphere.*

In particular, as indicated above, Cannon's conjecture is true for those hyperbolic groups whose boundary is quasi-Möbius homeomorphic to a Loewner 2-sphere. The higher dimensional analog of Theorem 6.8 is false, because the Carnot metric on S^3 is 4-Loewner but not quasi-Möbius homeomorphic to S^3.

Quasi-Möbius characterizations of the 2-sphere. In [5], the proofs of Theorem 6.8 and Conjecture 6.7 invoked a more general necessary and sufficient condition. To formulate this, we require a combinatorial version of modulus, and the related definitions.

Suppose $\mathcal{G}$ is a graph with vertex set $V = V(\mathcal{G})$, and Γ is a collection of subsets of V; in our context, the elements $\gamma \in \Gamma$ will be the vertex sets of certain connected subgraphs of Γ. A function $\rho\colon V(\mathcal{G}) \to [0, \infty)$ is Γ-*admissible* if for every $\gamma \in \Gamma$,

$$\sum_{v \in \gamma} \rho(v) \geq 1. \tag{6.4}$$

If $Q \geq 1$, the Q-*modulus of* Γ is the infimum of

$$\sum_{v \in \mathrm{Vertex}(\mathcal{G})} \rho^Q(v) \tag{6.5}$$

where ρ ranges over all Γ-admissible functions.

Pick $Q \geq 1$. Now suppose Γ is a path family in a metric space Z, and $\mathcal{U}$ is an open cover of Z. Then the Q-*modulus of* Γ *with respect to* $\mathcal{U}$, denoted $\mathrm{Mod}_Q(\Gamma, \mathcal{U})$, is defined as follows. We let $\mathcal{G}(\mathcal{U})$ be the nerve of the open cover $\mathcal{U}$, which is the graph with vertex set $\mathcal{U}$ whose edges correspond to pairs $U, U' \in \mathcal{U}$ with nonempty intersection. Then we let $\Gamma(\mathcal{G})$ be the collection of subsets of the vertex set $V(\mathcal{G})$ of the form

$$\{U \in \mathcal{U} \mid U \cap \mathrm{Im}\, \gamma \neq \emptyset\},$$

where γ ranges over all paths $\gamma \in \Gamma$. Note that each element of $\Gamma(\mathcal{G})$ is the 0-skeleton of a connected subgraph if $\mathcal{G}$. We then define $\mathrm{Mod}_Q(\Gamma, \mathcal{U})$ to be the Q-modulus of $\Gamma(\mathcal{G})$ in $\mathcal{G}$. Finally, if $E, F \subset Z$ are subsets, we let $\mathrm{Mod}_Q(E, F; \mathcal{U}) := \mathrm{Mod}_Q(\Gamma(E, F), \mathcal{U})$, where $\Gamma(E, F)$ denotes the family of paths joining E to F.

Theorem 6.9 ([5]). *Let Z be a doubling, linearly locally contractible metric 2-sphere, and let $\{r_i\}$ be a sequence of positive numbers converging to 0. For each i, let V_i be a maximal r_i-separated subset of Z, and let*

$$\mathcal{U}_i := \{B(v, r_i)\}_{v \in V_i} \tag{6.6}$$

be the corresponding open ball cover. Then the following conditions are equivalent.

- *Z is quasi-Möbius homeomorphic to $\mathbb{S}^2$.*
- *There is a function $\psi\colon [0, \infty) \to [0, \infty]$ tending to zero at infinity, and a number L, such that if $E, F \subset Z$ are closed subsets, then*

$$\mathrm{Mod}_2(E, F; \mathcal{U}_i) \leq \psi(\Delta(E, F)) \tag{6.7}$$

for every i satisfying

$$\min(\mathrm{diam}(E), \mathrm{diam}(F)) \geq L\, r_i.$$

Here $\Delta(E, F)$ denotes the relative distance as before.

- *There is a positive decreasing function* $\phi\colon [0,\infty) \to (0,\infty)$ *and a number* M *such that if* $E, F \subset Z$ *are continua, then*

$$\mathrm{Mod}_2(E, F; \mathcal{U}_i) \geq \phi(\Delta(E, F)), \tag{6.8}$$

for every i *satisfying*

$$\mathrm{dist}(E, F) \geq M\, r_i.$$

The theorem says that Z is quasi-Möbius equivalent to the standard 2-sphere if and only if, for a sequence of combinatorial approximations, the combinatorial 2-modulus behaves as in $\mathbb{S}^2$, i.e. it can be bounded below or above by functions of relative distance. The idea of the proof is to associate, for each i, a topological triangulation $\mathcal{T}_i$ of Z whose 1-skeleton is quasi-isometric to the nerve of $\mathcal{U}_i$ (with quasi-isometry constants independent of i). Then one can apply classical uniformization to the equilateral polyhedron associated with $\mathcal{T}_i$ to produce a map $f_i\colon V_i \to \mathbb{S}^2$. The crux of the argument is to show that the maps f_i, when appropriately normalized, are uniformly quasi-Möbius, and hence subconverge to a quasi-Möbius homeomorphism by the Arzela–Ascoli theorem.

Other uniformization problems. One may formulate uniformization problems for spaces other than spheres. The case of Sierpinski carpets is especially interesting, where there are remarkable uniformization and rigidity results. We refer the reader to Mario Bonk's article in these Proceedings for a treatment of this topic.

7. Geometrization

Minimizing Hausdorff dimension. Geometrization – the problem of finding optimal or canonical geometric structures – appears in many contexts in mathematics: for example, the Yamabe problem (finding conformally equivalent metrics of constant scalar curvature), Thurston's Geometrization conjecture for 3-manifolds, Thurston's characterization of rational maps, and the Calabi conjecture. For each of these problems there are cases where there is no solution: the Yamabe problem for the "teardrop" 2-orbifold has no solution, and any closed 3-manifold whose prime or torus decomposition is nontrivial does not admit a Thurston geometry. The goal is to show that geometrization is always possible unless some alternate structure appears, on which the failure can be blamed (such as a bad conformal group in the case of the teardrop, or an essential decomposition in the 3-manifold geometrization problem).

In the metric space context, a natural geometrization problem is to minimize the Hausdorff dimension in an attempt to optimize shape. This leads to the following notion, which is a minor variant of a definition of Pansu:

Definition 7.1. The *conformal dimension of a metric space* Z is the infimal Hausdorff dimension of the Ahlfors regular metric spaces quasi-Möbius homeomorphic to it. We denote this by $\mathrm{Confdim}(Z)$.

Visual metrics on boundaries of hyperbolic groups are Ahlfors regular, so

$$\mathrm{Confdim}(\partial G) < \infty$$

for every hyperbolic group G. Likewise

$$\mathrm{Confdim}(Z) < \infty$$

for every self-similar space Z. Every Q-Loewner metric space (Definition 4.5), and more generally Q-regular metric spaces with nontrivial Q-modulus, are solutions to the geometrization problem, because they minimize Hausdorff dimension in their quasi-Möbius homeomorphism class:

Theorem 7.2 (Bonk–Tyson, see [37, Theorem 15.10]). *Suppose Z is a compact Ahlfors Q-regular metric space which carries a family of nonconstant paths of positive Q-modulus. Then any metric space quasi-Möbius homeomorphic to Z has positive Q-dimensional Hausdorff measure, and in particular $Q = \mathrm{Confdim}(Z)$.*

The following theorem of Keith–Laakso shows that the converse is nearly true:

Theorem 7.3 ([42]). *If Z is an Ahlfors Q-regular metric space where*

$$Q = \mathrm{Confdim}(Z) > 1,$$

then some weak tangent of Z carries a nontrivial curve family of positive Q-modulus.

Here a *weak tangent* is a pointed Gromov–Hausdorff limit of a sequence of rescalings of Z, see [42], [6]. If one adds the hypothesis that Z is self-similar, or approximately self-similar like a visual metric on the boundary of a hyperbolic group, then one can map open subsets of a weak tangent to Z itself, and thereby conclude that Z itself has a nontrivial curve family of positive Q-modulus. Therefore Theorem 7.2 has a converse in the self-similar case. The idea of the proof of Theorem 7.3 is to show that if the conclusion of the theorem fails for some Q-regular space Z, then one can use a construction of Semmes (a "Semmes deformation") to produce an Ahlfors regular metric of strictly smaller Hausdorff dimension.

For spaces quasi-Möbius homeomorphic to boundaries of groups, a much stronger statement holds:

Theorem 7.4 ([6]). *If Z is an Ahlfors Q-regular metric space where*

$$Q = \mathrm{Confdim}(Z) > 1,$$

and Z is quasi-Möbius homeomorphic to the boundary of some hyperbolic group G, then Z is Q-Loewner.

The proof of this theorem uses Theorem 7.3, work of Tyson [61], and a dynamical argument to show that for a large supply of ball pairs $B \subset B'$ the pair $(B, \overline{Z \setminus B'})$ has Q-modulus bounded away from zero. The main work consists in showing that this "ball-Loewner" condition implies the usual Loewner property in Definition 4.5.

Theorem 7.4 connects the problem of realizing the conformal dimension with results such as Theorems 5.6 and 6.8, since one would like to know for which hyperbolic groups the boundary is quasi-Möbius homeomorphic to a Loewner space.

Examples where the conformal dimension is (not) realized. Suppose G is an infinite hyperbolic group and $Q = \mathrm{Confdim}(\partial G)$ can be realized by an Ahlfors Q-regular metric. If $Q < 1$ then one can argue that G must be virtually infinite cyclic. If $Q = 1$, then it is not difficult to deduce that ∂G is homeomorphic to a circle, and hence by [23], [28] it is virtually a surface group. If $Q > 1$, then Theorem 7.4 implies that ∂G is quasi-Möbius homeomorphic to a Loewner space, which implies that it is connected and has no local cut points. These two conditions are equivalent by [4], [15], [58], [14] to saying that G does not virtually split over a virtually cyclic group.

Any hyperbolic group G which splits over a finite group has disconnected boundary, so unless G is virtually cyclic, the previous paragraph implies that the conformal dimension of ∂G cannot be realized. A free group of rank at least two is such an example.

Pansu showed that if one takes two copies of a surface of genus 2 and glues them along a homotopically nontrivial simple closed curve, the fundamental group of the resulting 2-complex is a hyperbolic group G where $\mathrm{Confdim}(\partial G) = 1$. Since G is not a virtual surface group, this group provides another example where the conformal dimension cannot be realized. In the case of self-similar spaces (not arising as boundaries of group), it was shown by Laakso that the Sierpinski gasket has conformal dimension 1, but it cannot be realized.

Much deeper examples were constructed by Bourdon and Pajot [13], [10]. These are hyperbolic groups which do not virtually split over virtually cyclic groups, and whose boundaries are not quasi-Möbius homeomorphic to Loewner spaces. These examples are very intriguing, and have led Marc Bourdon to speculate that the nonexistence of Loewner structure implies the presence of fibration-like structure in ∂G which is invariant under the group of quasisymmetric homeomorphisms.

A further necessary condition for the conformal dimension to be realized. Suppose Z is a compact doubling metric space, and $Q > 1$. Then Z satisfies the *combinatorial Q-Loewner property* if the following combinatorial analog of (4.6) and (4.8) holds. There are functions $\phi\colon [0, \infty) \to (0, \infty)$, $\psi\colon [0, \infty) \to [0, \infty]$ and a constant $\lambda > 0$, with the following properties:

- ϕ is a positive decreasing function, $\psi(t) \to 0$ as $t \to \infty$, and $\phi \leq \psi$.
- For every $0 < r \leq \mathrm{diam}(Z)$, if V is a maximal r-separated net in Z, $\mathcal{U}$ is the corresponding r-ball cover, and $E, F \subset Z$ are disjoint nondegenerate continua where $r \leq \lambda \min(\mathrm{diam}(E), \mathrm{diam}(F))$, then

$$\phi(\Delta(E, F)) \leq \mathrm{Mod}_Q(E, F; \mathcal{U}) \leq \psi(\Delta(E, F)), \tag{7.1}$$

where $\mathrm{Mod}_Q(E, F, \mathcal{U})$ is the combinatorial modulus defined just before Theorem 6.9.

Using arguments from [38], [5] is not hard to see that if Z is a Q-Loewner space, then Z satisfies the combinatorial Q-Loewner property. Based on current evidence, the following seems plausible:

Conjecture 7.5. Suppose Z is a self-similar space, or quasi-Möbius homeomorphic to the boundary of a hyperbolic group. If Z satisfies the combinatorial Loewner property, then Z is quasi-Möbius homeomorphic to a Loewner space.

Currently there is no example of a compact doubling space satisfying the combinatorial Loewner property, which is known not to be quasi-Möbius homeomorphic to a Loewner space.

The conjecture is intriguing, because the author has shown that several examples, including the standard square Sierpinski carpet, the standard Menger sponge (obtained from the unit cube in $\mathbb{R}^3$), and boundaries of certain hyperbolic Coxeter groups satisfy the combinatorial Loewner property. Therefore the conjecture would provide new examples of Loewner spaces.

8. Open problems

We conclude with some open problems. These are questions which seem to be key to making further progress with the central themes of this article.

Question 8.1. Let Z be the boundary of a hyperbolic group, or more generally an "approximately self-similar" space. When is the conformal dimension of Z realized?

Question 8.2. Suppose G is a Gromov hyperbolic group. Suppose G does not virtually split over a virtually cyclic group, or equivalently, that ∂G is connected and has no local cut points. Is every quasiconformal homeomorphism of ∂G quasi-Möbius?

Question 8.3. Is the standard square Sierpinski carpet quasi-Möbius homeomorphic to a Loewner space? What is its conformal dimension?

The author and independently [42] have shown that usual metric does not realize the conformal dimension. The author has shown that the square carpet satisfies the combinatorial Q-Loewner property, where Q is its conformal dimension, see Section 7.

Question 8.4. If G is a random hyperbolic group, is the homomorphism $G \to \mathrm{QI}(G)$ an isomorphism?

See [24], [41] for discussion of random groups.

Question 8.5. What are the quasi-isometry groups of the Gromov–Thurston examples [36]?

References

[1] Bass, H., The degree of polynomial growth of finitely generated nilpotent groups. *Proc. London Math. Soc.* (3) **25** (1972), 603–614.

[2] Besson, G., Courtois, G., and Gallot, S., Volumes, entropies et rigidités des espaces localement symétriques de courbure strictement négative. *C. R. Acad. Sci. Paris Sér. I Math.* **319** (1) (1994), 81–84.

[3] Bestvina, M., Local homology properties of boundaries of groups. *Michigan Math. J.* **43** (1) (1996), 123–139.

[4] Bestvina, M., and Mess, G., The boundary of negatively curved groups. *J. Amer. Math. Soc.* **4** (3) (1991), 469–481.

[5] Bonk, M., and Kleiner, B., Quasisymmetric parametrizations of two-dimensional metric spheres. *Invent. Math.* **150** (1) (2002), 127–183.

[6] Bonk, M., and Kleiner, B., Conformal dimension and Gromov hyperbolic groups with 2-sphere boundary. *Geom. Topol.* **9** (2005), 219–246.

[7] Bonk, M., and Kleiner, B., Quasi-hyperbolic planes in hyperbolic groups. *Proc. Amer. Math. Soc.* **133** (9) (2005), 2491–2494.

[8] Bourdon, M., Immeubles hyperboliques, dimension conforme et rigidité de Mostow. *Geom. Funct. Anal.* **7** (2) (1997), 245–268.

[9] Bourdon, M., Sur les immeubles fuchsiens et leur type de quasi-isométrie. *Ergodic Theory Dynam. Systems* **20** (2) (2000), 343–364.

[10] Bourdon, M., Cohomologie l_p et produits amalgamés. *Geom. Dedicata* **107** (2004), 85–98.

[11] Bourdon, M., and Pajot, H., Poincaré inequalities and quasiconformal structure on the boundary of some hyperbolic buildings. *Proc. Amer. Math. Soc.* **127** (8) (1999), 2315–2324.

[12] Bourdon, M., and Pajot, H., Rigidity of quasi-isometries for some hyperbolic buildings. *Comment. Math. Helv.* **75** (4) (2000), 701–736.

[13] Bourdon, M., and Pajot, H., Cohomologie l_p et espaces de Besov. *J. Reine Angew. Math.* **558** (2003), 85–108.

[14] Bowditch, B. H., Cut points and canonical splittings of hyperbolic groups. *Acta Math.* **180** (2) (1998), 145–186.

[15] Bowditch, B. H., Connectedness properties of limit sets. *Trans. Amer. Math. Soc.* **351** (9) (1999), 3673–3686.

[16] Bridson, M. R., and Haefliger, A., *Metric spaces of non-positive curvature*. Grundlehren Math. Wiss. 319, Springer-Verlag, Berlin 1999.

[17] Cannon, J. W., The combinatorial Riemann mapping theorem. *Acta Math.* **173** (2) (1994), 155–234.

[18] Cannon, J. W., Floyd, W. J., Kenyon, R., and Parry, W. R., Constructing rational maps from subdivision rules. *Conform. Geom. Dyn.* **7** (2003), 76–102.

[19] Cannon, J. W., Floyd, W. J., and Parry, W. R., Conformal modulus: the graph paper invariant or the conformal shape of an algorithm. In *Geometric group theory down under* (Canberra, 1996), Walter de Gruyter, Berlin 1999, 71–102.

[20] Cannon, J. W., Floyd, W. J., and Parry, W. R., Sufficiently rich families of planar rings. *Ann. Acad. Sci. Fenn. Math.* **24** (2) (1999), 265–304.

[21] Cannon, J. W., Floyd, W. J., and Parry, W. R., Finite subdivision rules. *Conform. Geom. Dyn.* **5** (2001), 153–196.

[22] Cannon, J. W., and Swenson, E. L., Recognizing constant curvature discrete groups in dimension 3. *Trans. Amer. Math. Soc.* **350** (2) (1998), 809–849.

[23] Casson, A., and Jungreis, D., Convergence groups and Seifert fibered 3-manifolds. *Invent. Math.* **118** (3) (1994), 441–456.

[24] Champetier, C., Propriétés statistiques des groupes de présentation finie. *Adv. Math.* **116** (2) (1995), 197–262.

[25] Cheeger, J., Differentiability of Lipschitz functions on metric measure spaces. *Geom. Funct. Anal.* **9** (3) (1999), 428–517.

[26] Coornaert, M., Mesures de Patterson-Sullivan sur le bord d'un espace hyperbolique au sens de Gromov. *Pacific J. Math.* **159** (1993), 241–270.

[27] Corlette, K., Archimedean superrigidity and hyperbolic geometry. *Ann. of Math.* (2) **135** (1) (1992), 165–182.

[28] Gabai, D., Convergence groups are Fuchsian groups. *Ann. of Math.* (2) **136** (3) (1992), 447–510.

[29] Gabai, D., Meyerhoff, G. R., and Thurston, N., Homotopy hyperbolic 3-manifolds are hyperbolic. *Ann. of Math.* (2) **157** (2) (2003), 335–431.

[30] Ghys, É., and de la Harpe, P. (eds.), *Sur les groupes hyperboliques d'après Mikhael Gromov*. Progr. Math. 83 Birkhäuser, Boston, MA, 1990.

[31] Gromov, M., Hyperbolic groups. In *Essays in group theory*, Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–263.

[32] Gromov, M., Hyperbolic manifolds, groups and actions. In *Riemann surfaces and related topics: Proceedings of the 1978 Stony Brook Conference* (State Univ. New York, Stony Brook, N.Y., 1978), Ann. of Math. Stud. 97, Princeton University Press, Princeton, N.J., 1981, 183–213.

[33] Gromov, M., Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[34] Gromov, M., Asymptotic invariants of infinite groups. In *Geometric group theory* (Sussex, 1991), Vol. 2, Cambridge University Press, Cambridge 1993, 1–295.

[35] Gromov, M., and Schoen, R., Harmonic maps into singular spaces and p-adic superrigidity for lattices in groups of rank one. *Inst. Hautes Études Sci. Publ. Math.* (76) (1992), 165–246.

[36] Gromov, M., and Thurston, W., Pinching constants for hyperbolic manifolds. *Invent. Math.* **89** (1) (1987), 1–12.

[37] Heinonen, J., *Lectures on analysis on metric spaces*. Universitext, Springer-Verlag, New York 2001.

[38] Heinonen, J., and Koskela, P., Quasiconformal maps in metric spaces with controlled geometry. *Acta Math.* **181** (1) (1998), 1–61.

[39] Heinonen, J., Koskela, P., Shanmugalingam, N., and Tyson, J. T., Sobolev classes of Banach space-valued functions and quasiconformal mappings. *J. Anal. Math.* **85** (2001), 87–139.

[40] Heinonen, J., and Semmes, S., Thirty-three yes or no questions about mappings, measures, and metrics. *Conform. Geom. Dyn.* **1** (1997), 1–12.

[41] Kapovich, I., and Benakli, N., Boundaries of hyperbolic groups. In *Combinatorial and geometric group theory* (New York, 2000/Hoboken, NJ, 2001), Contemp. Math. 296, Amer. Math. Soc., Providende, RI, 39–93.

[42] Keith, S., and Laakso, T., Conformal Assouad dimension and modules. *Geom. Funct. Anal.* **14** (6) (2004), 1278–1321.

[43] Kleiner, B., In preparation, 2005.

[44] Kleiner, B., and Leeb, B., Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings. *Inst. Hautes Études Sci. Publ. Math.* **86** (1997), 115–197.

[45] Korányi, A., and Reimann, H. M., Quasiconformal mappings on the Heisenberg group. *Invent. Math.* **80** (2) (1985), 309–338.

[46] Laakso, T. J., Plane with A_∞-weighted metric not bi-Lipschitz embeddable to $\mathbb{R}^N$. *Bull. London Math. Soc.* **34** (6) (2002), 667–676.

[47] Margulis, G. A., Arithmeticity of the irreducible lattices in the semisimple groups of rank greater than 1. *Invent. Math.* **76** (1) (1984), 93–120.

[48] Margulis, G. A., and Mostow., G. D., The differential of a quasi-conformal mapping of a Carnot-Carathéodory space. *Geom. Funct. Anal.* **5** (2) (1995), 402–433.

[49] Mostow., G. D., *Strong rigidity of locally symmetric spaces*. Ann. of Math. Stud. 78, Princeton University Press, Princeton, N.J., 1973.

[50] Pansu, P., Croissance des boules et des géodésiques fermées dans les nilvariétés. *Ergodic Theory Dynam. Systems* **3** (3) (1983), 415–445.

[51] Pansu, P., Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rang un. *Ann. of Math.* (2) **129** (1) (1989), 1–60.

[52] Paulin, F., Un groupe hyperbolique est déterminé par son bord. *J. London Math. Soc.* (2) **54** (1) (1996), 50–74.

[53] Prasad, G., Strong rigidity of **Q**-rank 1 lattices. *Invent. Math.* **21** (1973), 255–286.

[54] Semmes, S., Good metric spaces without good parameterizations. *Rev. Mat. Iberoamericana* **12** (1) (1996), 187–275.

[55] Shalom, Y., Harmonic analysis, cohomology, and the large-scale geometry of amenable groups. *Acta Math.* **192** (2) (2004), 119–185.

[56] Stallings, J. R., On torsion-free groups with infinitely many ends. *Ann. of Math.* (2) **88** (1968), 312–334.

[57] Sullivan, D., Discrete conformal groups and measurable dynamics. *Bull. Amer. Math. Soc.* (*N.S.*) **6** (1) (1982), 57–73.

[58] Swarup, G. A., On the cut point conjecture. *Electron. Res. Announc. Amer. Math. Soc.* **2** (2) (1996), 98–100.

[59] Tukia, P., Homeomorphic conjugates of Fuchsian groups. *J. Reine Angew. Math.* **391** (1988), 1–54.

[60] Tukia, P., and Väisälä, J., Quasisymmetric embeddings of metric spaces. *Ann. Acad. Sci. Fenn. Ser. A I Math.* **5** (1) (1980), 97–114.

[61] Tyson, J., Quasiconformality and quasisymmetry in metric measure spaces. *Ann. Acad. Sci. Fenn. Math.* **23** (2) (1998), 525–548.

[62] Väisälä, J., *Lectures on n-dimensional quasiconformal mappings*. Lecture Notes in Math. 229, Springer-Verlag, Berlin 1971.

[63] Väisälä, J., Quasi-Möbius maps. *J. Analyse Math.* **44** (1984/85), 218–234.

[64] Xie, X., Quasi-isometric rigidity of Fuchsian buildings. *Topology* **45** (1) (2006), 101–169.

Mathematics Department, Yale University, New Haven, CT 06511, U.S.A.
E-mail: bruce.kleiner@yale.edu

Lagrangian submanifolds: from the local model to the cluster complex

François Lalonde

Abstract. In these notes, I will present collaborative works on Lagrangian submanifolds that I realised mainly with Octav Cornea (the cluster complex, which naturally leads to a universal Lagrangian Floer theory), but also, at an earlier stage, with Jean-Claude Sikorav. To cover the subject in a more complete and adequate way, I will also mention very recent works by Barraud–Cornea and by Welschinger, closely related to the subject of these notes. The aim of the cluster machinery is to resolve the well known problem of real codimension 1 bubbling off of disks in the Gromov–Floer theory; see Fukaya–Oh–Ohta–Ono (especially the two lectures by Oh and Ono in these proceedings) for a different, earlier, approach.

Mathematics Subject Classification (2000). Primary 53D12; Secondary 53D40.

Keywords. Lagrangian submanifold, cluster homology, Floer complex, J-holomorphic curves, Morse theory.

1. Introduction

By Weinstein's theorem, a sufficiently small neighbourhood of any Lagrangian submanifold $L \subset M$ of a symplectic manifold is symplectomorphic to a neighbourhood of the zero section of the cotangent T^*L of L through a diffeomorphism that sends L to the zero section. The study of Lagrangian submanifolds, which play a fundamental role in our understanding of symplectic geometry, therefore breaks down into two parts:

Local model: study of exact Lagrangian submanifolds in cotangent spaces.

General case: study of Lagrangian submanifolds in general symplectic manifolds.

From a heuristic point of view, these two cases correspond to the following dichotomy in studying Lagrangian submanifolds: either we are in a case where "real bubbling off" of disks is prohibited or not. The former case is much simpler, and the simplest case in which bubbling off cannot occur is the one of closed embedded exact Lagrangian submanifolds in cotangent spaces $L \subset T^*V$ where V is any (not necessarily closed) manifold and where T^*V is equipped with the exact symplectic form $\omega = d\lambda$, with λ the Liouville form on T^*V. Recall that λ is equal to the form $\sum_i p_i dq_i$ in local coordinates where the q_i's are coordinates on V and the p_i's the coordinates naturally induced on the fibers of the cotangent bundle by the q_i's. A

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

submanifold $L \subset T^*V$ is *Lagrangian* if the restriction of ω to L vanishes identically, and it is *exact* if the restriction of λ (which is therefore closed on L) is exact on L. In a general symplectic manifold (M, ω), the notion of a Lagrangian submanifold is defined similarly, and the exactness is replaced by the notion of *weakly exact* meaning that the restriction of ω to $\pi_2(M, L)$ vanishes. It is obvious, by Stokes' Theorem, that an exact Lagrangian submanifold is weakly exact. However, as far as analytic techniques are concerned, there is not much difference between the two concepts (although the geometry might be quite different). See [10] for the definition of (generic) almost complex structures J's on a symplectic manifold tamed by ω. By definition, for such J's, J-invariant real 2-planes in the tangent space of any symplectic manifold are ω-positive – thus non-constant J-pseudoholomorphic curves, also called J-curves, (i.e. those for which the real differential commutes with the complex structure i on the Riemann surface at the source and the J-structure at the target space) have strictly positive symplectic area. *Real bubbling off* is the phenomenon that occurs when a J-curve with boundaries on a finite set of Lagrangian submanifolds of a symplectic manifold degenerates to the connected union of a J-curve connecting the same set of Lagrangian submanifolds and a J-pseudoholomorphic disk with boundary on one of the submanifolds. Thus real bubbling off can only happen if at least one of the classes in $\pi_2(M, L_i)$, for at least one i, has strictly positive symplectic area. Hence there is no bubbling off in (weakly) exact Lagrangian submanifolds, which makes the analytic study much simpler. See [2] for a survey on Lagrangian submanifolds.

2. Exact Lagrangian submanifolds

The most obvious examples of Lagrangian (exact) submanifolds are the graphs, in T^*V, of closed (exact) 1-forms defined on V. It turns out that exact Lagrangian submanifolds behave in a much more rigid way than other Lagrangian submanifolds, so much that the following statement seems a reasonable (although very strong) conjecture:

Conjecture A (see Lalonde–Sikorav [14]). Any closed exact Lagrangian submanifold of a cotangent space is isotopic, through a Hamiltonian isotopy, to the zero section.

A *Hamiltonian isotopy* is a very constraining condition: it requires that the isotopy be Lagrangian at each moment and that the symplectic area of the cylinder spanned by any non-contractible loop during any time interval in the isotopy be zero. Note that this conjecture implies that there is no closed exact Lagrangian submanifold in the cotangent of an open manifold, a fact proved in [14]. Actually, we proved in [14] that the projection $L \to V$ must be surjective in the exact case. This is a consequence of the following "Property 4" of [14]:

Let $K \subset V$ be a closed manifold. Then any closed exact Lagrangian submanifold L must intersect the conormal νK of K in the following two cases: (i) *K is the*

fiber of a submersion $V \to B$ where B is a closed submanifold; (ii) *K is homotopic to a point in V.*

Since the conormal of a point $q \in V$ is the fiber $T^*_q V$ at that point, surjectivity follows from (ii) (see below for a more direct proof). From (i), one may deduce that any two closed exact Lagrangian submanifolds L_0, L_1 must intersect in the cotangent of any homogeneous space $V = G/H$. Indeed, by fiber product, one easily assigns to L_0, L_1 two closed exact Lagrangian submanifolds L'_0, L'_1 of T^*G with a natural bijection between $L_0 \cap L_1$ and $L'_0 \cap L'_1$. But the latter is in bijection with

$$(L'_0 \times L'_1) \cap \Delta_{T^*G} \subset T^*G \times T^*G$$

or, equivalently, with $(L'_0 \times -L'_1) \cap A$ where $A \subset T^*G \times T^*G$ is the anti-diagonal. But $T^*G \times T^*G = T^*(G \times G)$, A is the conormal of Δ_G which is a fiber of the submersion $(g, h) \mapsto gh^{-1}$, so the fact that $(L'_0 \times -L'_1) \cap A$ is not empty is a consequence of Property 4 (i) above.

Conjecture A has many other consequences, that could happen to be true while the conjecture itself might turn out to be false. Denoting by L a closed exact Lagrangian submanifold of T^*V, here are some of these consequences:

(I) L is diffeomorphic to V.

(II) The Maslov index of L vanishes.

(III) The map induced at the π_1-level by the projection $L \to V$ is onto.

(IV) The projection $L \to V$ has degree ± 1.

Except in cotangent spaces of surfaces ($n = 2$), the first of these consequences seems out of reach for the moment. Actually, in the case $n = 2$, many of these properties have been solved and all of them have been established for the torus. Indeed, it is not difficult to see that the formula for the number of double points of an immersed Lagrangian submanifold must be

$$\frac{d^2\chi(V) - \chi(L)}{2}$$

in the orientable case, which implies that the formula $\chi(L) = d^2\chi(V)$ must hold for an embedded Lagrangian submanifold (here d is the degree of the projection $L \to V$). Hence, in dimension 2, (I) is a consequence of (IV) (a similar formula mod 2 holds in the non-orientable case as well). It turns out that (IV) was established for T^2 in [14], and that, for surfaces, the degree $L \to V$ was proved to be always non-zero with the sole exception of an eventual exact degree 0 Lagrangian embedding of the 2-torus in the cotangent of the 2-sphere, a case ruled out by Viterbo a few years later (see [23]) using Floer techniques (or equivalently techniques related to the cohomology of the free loop space).

Let me now briefly explain, from [14], how one can prove all of these consequences, except the first one, in the cotangent of the n-torus, starting from the following basic results proved by Gromov using pseudoholomorphic curves:

Theorem 1 (Gromov [10]). (1) *A closed exact Lagrangian submanifold must intersect the zero section.*

(2) *A closed exact Lagrangian submanifold must intersect any submanifold obtained as the image of itself by a Hamiltonian isotopy.*

(The second statement in Gromov's theorem has been pushed much further by Floer, but this is not relevant at this moment).

We will also use Audin's conjecture, whose proof has recently been announced by several authors, stating that any Lagrangian embedding $L = T^n \to \mathbb{R}^{2n}$ has Maslov number 2 (i.e. there is a loop on L whose Maslov index is 2). Actually, we do not need such a strong conjecture for our present purposes: any bound, say the one found by Viterbo in [22], stating that there is a loop with Maslov index in $[2, n+1]$ for any such embedding, is enough.

Here is how to prove (III) for a closed exact Lagrangian submanifold L in $M = T^*T^n$. First note that the index of $\pi_\#(\pi_1(L))$ in $\pi_1(V)$ is always finite: otherwise, denoting by $V_1 \to V$ the covering corresponding to $\pi_\#(\pi_1(L))$, one could then lift L to a Lagrangian embedding $L_1 \subset T^*V_1$. If the index were infinite, one would then get a closed exact Lagrangian submanifold in the cotangent of an open manifold. The image of the projection $L_1 \to V_1$ would not be surjective. But this is impossible: indeed, if it were not onto, one could then define a Morse function f on V_1 with all of its critical points outside the subset $\pi(L_1)$ of V_1. Translating L_1 by a sufficiently large multiple of df in the fibers of T^*V_1 would produce an exact Lagrangian submanifold that would no longer intersect the zero section, a contradiction with Theorem 1 (1) above.

Now let us show that $f_\#\colon \pi_1(L) \to \pi_1(V)$ must actually be surjective when V is a Lie group. This will prove (III) (and therefore (IV) when V is a torus, if we know that L is also a n-torus). The basic geometric idea is that, if this map were not surjective, one would get more than one disjoint lifts of the same Lagrangian submanifold in the cotangent space of some covering V' of V corresponding to the subgroup $\pi_\#(\pi_1(L))$ in $\pi_1(V)$. But the deck transformations of that covering are induced by translations of the Lie group, and are therefore isotopic to the identity. Thus the group of automorphisms, at the cotangent level, of $T^*V' \to T^*V$, consists of Hamiltonian diffeomorphisms (this is because the differential of any diffeomorphism ϕ of a manifold V' induces a symplectic diffeomorphism of T^*V' and one sees easily that this symplectomorphism is Hamiltonian isotopic to the identity if ϕ is isotopic to the identity). This shows that one gets many disjoint exact Lagrangian embeddings of L, all Hamiltonian isotopic to each other, a contradiction with Theorem 1 (2).

Finally, the second consequence (the vanishing of the Maslov class) for an exact n-torus in the cotangent of the n-torus is proved using the geometric composition $j \circ i$ of two Lagrangian embeddings $i\colon L \to T^*V$ and $j\colon V \to (\mathbb{R}^{2n}, \omega_0)$ given by extending j to a small neighbourhood of the zero section by Weinstein's theorem and then composing after contraction of i in the fibers. The Maslov class formula is $\mu(j \circ i) = \mu(i) + f^*(\mu(j))$ where f is the projection of L on V. One can

start with the standard embedding j and iterate the construction. By the surjectivity of $f_\#$ proved above, one deduces that $f^*\colon H^1(V,\mathbb{Z}) \to H^1(L,\mathbb{Z})$ is a bijection, which gives enough control to exceed any bound imposed by Audin's conjecture or by Viterbo's theorem if $\mu(i)$ does not vanish and if we iterate that construction a sufficient number of times (see [14] for more detail).

Using the Lagrangian surgery introduced in [14], and developed shortly afterwards by L. Polterovich, it was easy to say exactly in [14] which surfaces can be embedded in the cotangent of any surface (orientable or not), with the sole exception of an eventual local embedding of the Klein bottle, a difficult and classical problem whose solution has recently been explored by various methods, especially the ones of symplectic field theory.

2.1. Methods of symplectic field theory. There are, currently, two promising ways of establishing (some of) the above consequences (I)–(IV). The first one is symplectic field theory. The best results obtained so far are in real dimension 4 (cotangent of surfaces):

Let V be either the 2-sphere or the 2-torus. Then any closed orientable exact Lagrangian submanifold L in T^*V is Hamiltonian isotopic to the zero section.

To prove this, note that by Lalonde–Sikorav [14] and Viterbo [23] above, the degree cannot be zero which means, by the double point formula, that $\chi(L) = \chi(V)$; moreover, in the case $V = S^2$, the same formula implies that the degree is 1 while in the case $V = T^2$, this is Consequence (IV) proved above for tori. Thus L is diffeomorphic to V and is homologous to the zero section. On the other hand, building on a result by Eliashberg–Polterovich [7], Hind (preprint "Lagrangian unknottedness in Stein surfaces" based on [11]) and A. Ivrii [13] showed, using methods of SFT, that such a surface is Lagrangian isotopic to the zero section; since we have proved above that the projection in the torus case induces an isomorphism at the $H^1(\,\cdot\,;\mathbb{R})$-level, one may change the isotopy at each time by translating by the graph of an appropriate closed one-form in order to make the isotopy exact, which proves the statement. See the article by Eliashberg in these Proceedings for more on SFT.

2.2. A natural more algebraic approach. The second promising approach is implicit in the methods of [14]: one very simple idea that inspired the results of that paper is that if the projection $L \to V$ is a covering (without singularities), then the degree must be ± 1. The reason is that, if $|d|$ were larger than 1, the differences $x - y$ between all pairs of distinct points $x, y \in L \cap T^*_q V$ would form an exact Lagrangian submanifold $\mathcal{D}(L)$ of T^*V and the resulting submanifold would not meet the zero section, a contradiction. Unfortunately, this simple argument breaks down when singularities of the projection $L \to V$ occur, because some differences may vanish at a singularity. However, this is a frustrating problem since one should recognize the right pairs. A natural way to do this is to replace the difference of pairs by pseudo-holomorphic strips joining pairs of such points, with one boundary on L and the other on the fiber. What is needed is to update the paper [14] by replacing Gromov's theory

by Floer's theory. Such an approach has been suggested by Fukaya for tori (private communication, Ringberg, 1997) and there is some hope that it will eventually lead to proofs of some of the consequences (II), (III) or (IV).

3. The cluster complex

The most useful algebraic tool for studying Lagrangian submanifolds in general symplectic manifolds is Floer homology. The goal is to assign to a pair L_0, L_1 of Lagrangian submanifolds a homology $FH_*(L_0, L_1)$ invariant under Hamiltonian deformations of any of the L_i (thus it would also assign an invariant to a single L by taking L and any of its images by a Hamiltonian isotopy). Assuming L_0, L_1 closed and meeting each other transversally, the complex is generated over $\mathbb{Q}$ by the finite set $I = L_0 \cap L_1$ and the differential is given by $da = \sum_{b\in I,\lambda} \mathrm{Card}(\mathcal{M}(a, b, \lambda; J))be^{\lambda}$ where λ keeps track of the homotopy class of strips joining a to b (parametrised say by the closed unit disk with two point p_-, p_+ removed, such that the lower boundary be sent to L_0, the upper one to L_1, and so that the map converge to a at p_- and to b at p_+) and where $\mathcal{M}(a, b, \lambda; J)$ is the moduli space of J-holomorphic strips joining a to b, with finite symplectic area, in the class λ, after quotient by the real 1-dimensional group of reparametrizations. The cardinality is taking over $\mathbb{Q}$ and is declared to be zero whenever the expected dimension of $\mathcal{M}(a, b, \lambda; J)$ is different from 0.

It is well known that this scheme does not work in general because d^2 is not always zero. The obstruction occurs when there is real bubbling off either on L_0 or on L_1. The reason is that the way to prove that d^2a vanishes follows the same argument as in ordinary Morse homology: an element in d^2a corresponds to two strips u_1 from a to b, and u_2 from b to c, each one belonging to a moduli space of dimension zero. The J-holomorphic surgery at b removes the singularity at b and exhibits the broken configuration (u_1, u_2) as the boundary of a real one-dimensional moduli space of strips u joining a to c. As this manifold $\mathcal{M}$ is compact, it must have another end: if that end is of the form (u_1', u_2') joining a to c through b', one sees that this cancels out the term ce^{λ} in d^2a. However, unfortunately, $\mathcal{M}$ might have an end of a different nature if its one-parameter family of strips degenerates to a configuration made from one strip joining a to c in class $\lambda - \tau$ and one J-holomorphic disk with boundary on either of the L_i in class τ, and meeting the strip at one of its boundary points on L_i (which, generically, can be assumed to be distinct from a and c). In this case, the resulting broken configuration does not correspond anymore to an element of d^2a and the classical cancellation argument breaks down.

With Octav Cornea, we have recently proposed in [5] a solution to overcome this problem by introducing a broader algebra and larger moduli spaces, large enough so that the above undesirable broken configurations be realised as *interior* points in those larger moduli spaces. In order to carry out this programme, the first step consists in defining, for each of the L_i separately, a new complex, the *cluster complex* of L_i, leading to a well-defined homology assigned to each Lagrangian submanifold; the

second step consists in using these complexes as coefficient rings to define a universal Floer homology, that we called the *Fine Floer homology* of the pair (L_0, L_1). From an algebraic point of view, the cluster complex has similarities with Chekanov's contact homology. From the perspective of the objectives, there are obvious similarities with the Fukaya–Oh–Ohta–Ono [8] approach – they introduced a universal object attached to each Lagrangian submanifold using an A^∞-approach and employed it as a coefficient ring; however our approach is based on a different geometric idea, leading to different moduli spaces and the relations between the cluster and the FOOO settings are still unclear.

Here is a description of the cluster homology, the Fine Floer homologies, and some of their applications.

We shall assume here that all Lagrangian submanifolds are compact, connected, orientable, and relative spin (recall that a Lagrangian submanifold $L \subset (M, \omega)$ is relative spin if the second Stiefel–Whitney class of L admits an extension to $H^2(M; \mathbb{Z}/2)$; a set of Lagrangian submanifolds is relative spin if their second Stiefel–Whitney classes admit a common extension to $H^2(M; \mathbb{Z}/2)$). In the notation L for a Lagrangian submanifold we always implicitly include the choices of an orientation and of a relative spin structure (the same applies for a set of such submanifolds) as described in [8]. The ambient symplectic manifold (M^{2n}, ω) is supposed to be compact or, if not, it should be geometrically bounded, so that no sequence of J-curves with boundary lying on a set of compact Lagrangian submanifolds $L_1, \dots, L_\ell \subset M$ can escape to infinity. In fact, the Lagrangian submanifolds need not all be compact, as long as the above control on sequences of J-curves is ensured.

The cluster complex is associated to a triple formed of (1) a Lagrangian embedding $L^n \hookrightarrow (M, \omega)$, equipped with a choice of an orientation and of a relative spin structure, (2) a generic almost complex structure J on M compatible with the symplectic form ω, and (3) a pair (f, g) with $f\colon L \to \mathbb{R}$ a Morse function and g a Riemannian metric on L so that (f, g) is Morse–Smale. The two conditions implicit in the notation L (i.e. the orientation and the relative spin structure) are needed to orient the clustered moduli spaces – to be roughly described below – in a coherent way.

This complex is denoted by $\mathcal{C}\ell_*(L; J, (f, g))$ and, with $\mathrm{Crit}(f)$ denoting the set of critical points of f, we set

$$\mathcal{C}\ell_*(L; J, (f, g)) = (S\mathbb{Q}\langle s^{-1}\,\mathrm{Crit}(f)\rangle \otimes\ \Lambda)^\wedge_*$$

where $s^{-1}\,\mathrm{Crit}(f)$ indicates that the natural index grading of $\mathrm{Crit}(f)$ is decreased by one unit, SV is the free, graded commutative algebra over the graded vector space V (as usual, the sign commutativity rule is $ab = (-1)^{|a||b|}ba$ for any two elements $a, b \in V$), Λ is the rational group ring of the quotient $\pi_2(M, L)/\sim$ where the equivalence relation $\sim$ is given by $\lambda \sim \tau$ iff $\omega(\lambda) = \omega(\tau)$ and $\mu(\lambda) = \mu(\tau)$, where ω and μ are the area and Maslov classes respectively. We write the elements of Λ in the form of finite sums $\sum_i c_i e^{\lambda_i}$, $c_i \in \mathbb{Q}$. The grading in Λ is given by $|e^\lambda| = -\mu(\lambda)$ for $\lambda \in \pi_2(M, L)/\sim$. With this convention, the grading of the cluster complex is given

by the usual tensor product formula. Thus for $x_i \in \mathrm{Crit}(f)$, we have

$$|x_i| = \mathrm{ind}_f(x_i) - 1, \quad |x_1 \dots x_k e^\lambda| = \sum_{i=1}^{k} |x_i| - \mu(\lambda).$$

Finally, we describe the completion $^\wedge$. An element $m \in \mathcal{C}\ell\,(L; J, (f, g))$ can be written as a possibly infinite sum

$$m = m_0 + m_1 e^{\lambda_1} + \cdots + m_i e^{\lambda_i} + \cdots$$

where m_i are monomials in the elements of $\mathrm{Crit}(f)$ but, if this sum is infinite, then any infinite subsequence with $\omega(\lambda_i)$ bounded above, must have its corresponding word length sequence converging to infinity. Conversely, any formal sum verifying this condition belongs to the cluster complex.

We now give a rough idea of the construction of the cluster differential. The generic data $J, (f, g)$ on L being given, fix an order on the critical points of f. Choose any integer $k \geq 0$, any $x \in \mathrm{Crit}(f)$, any non-decreasing sequence of critical points $x_1, \dots, x_k$, and $\lambda \in \pi_2(M, L)/\sim$, with the constraint that the zero class $\lambda = 0$ is allowed only when k equals 1. The cluster differential

$$d\colon (S\mathbb{Q}\langle s^{-1}\,\mathrm{Crit}(f)\rangle \otimes\ \Lambda)^\wedge_* \to (S\mathbb{Q}\langle s^{-1}\,\mathrm{Crit}(f)\rangle \otimes\ \Lambda)^\wedge_{*-1}$$

is the unique commutative, graded differential algebra extension of

$$dx = \sum_{\substack{\lambda, k \geq 0 \\ x_1, \dots, x_k}} a^x_{x_1,\dots,x_k}(\lambda) x_1 \dots x_k e^\lambda, \tag{1}$$

where $x, x_1, \dots, x_k \in \mathrm{Crit}(f)$ have the property that $(x_1, \dots, x_k)$ respects the fixed order on $\mathrm{Crit}(f)$, $|x| - \sum_i |x_i| + \mu(\lambda) - 1 = 0$ and the coefficients $a^x_{x_1,\dots,x_k}(\lambda)$ count with signs the number of elements in the clustered moduli spaces

$${}^{\nu}\mathcal{M}^x_{x_1,\dots,x_k}(\lambda)$$

(due to the possible presence of multi-sections this number will belong in fact to $\mathbb{Q}$).

A rigorous description of the clustered moduli spaces appears in [5]. The main idea in their definition is simple: consider a one parametric family of J-disks of class λ and assume that the bubbling off of a J-disk of class λ' occurs in this family. The "bubbled configuration" formed by two touching J-disks in classes $\lambda'' = \lambda - \lambda'$ and λ' can of course be viewed as the limit of this bubbling off, but it can also be considered as the limit of a one parametric family of two J-disks in the same classes $\lambda'' = \lambda - \lambda'$ and λ' joined by a negative gradient flowline of f whose length tends to 0. By gluing these two one-parameter moduli spaces at their common limit, the above bubbled configuration becomes an interior point of the larger clustered moduli space.

By pursuing systematically this idea together with the usual description of stable maps (as, for example, in McDuff–Salamon [15]) we obtain moduli spaces of configurations modelled on oriented trees with edges of non-negative length so that each vertex corresponds to a J-disk (or sphere), and each edge corresponds to a negative flow line of f joining incidence points situated on the boundaries of the disks whose corresponding vertices are related by that edge. Moreover, to define ${}^{\nu}\mathcal{M}^{x}_{x_1,\dots,x_k}(\lambda)$, such a configuration is supposed to also carry $k+1$ additional marked points $z, z_1, \dots, z_k$ so that z belongs to the boundary of the root disk (which corresponds to the root vertex of the tree) and the z_i's belong to the boundaries of some of the disks involved so that z is in the unstable manifold of the critical point x and z_i is in the stable manifold of x_i; the sum of the classes of the disks and spheres involved has to be λ. There are, of course, appropriate stability conditions and, to insure a reasonable structure for these moduli spaces, we need to use a system ν of perturbations of the pseudoholomorphic equation. The role of ghost disks (for which the corresponding J-disk is constant) is particularly important as they allow to deal not only with the transversality issues due to multiple coverings but also with the crossing of some of the incidence or marked points.

The dimension of ${}^{\nu}\mathcal{M}^{x}_{x_1,\dots,x_k}(\lambda)$ equals

$$|x| - \sum_{i=1}^{k} |x_i| + \mu(\lambda) - 1. \tag{2}$$

These moduli spaces admit natural compactifications ${}^{\nu}\overline{\mathcal{M}}^{x}_{x_1,\dots,x_k}(\lambda)$. To describe its properties, introduce the notation S for a partially ordered set of critical points (in which repetitions are allowed) which, of course, can be identified with a unique non-decreasing sequence of points. If S', S'' are two such ordered subsets, we will denote by $\langle S' \cup S'' \rangle$ the partially ordered subset made of the elements in $S' \cup S''$. Letting S be the ordered set $\{x_1, \dots, x_k\}$, for those moduli spaces of dimension 1 (which are, in fact, branched 1-dimensional manifolds with rational weights) we have:

$$\partial({}^{\nu}\overline{\mathcal{M}}^{x}_{S}(\lambda)) = \bigcup_{\substack{S=\langle S'\cup S''\rangle \\ y,\ \lambda'+\lambda''=\lambda}} ({}^{\nu}\overline{\mathcal{M}}^{x}_{\langle S',y\rangle}(\lambda')) \times ({}^{\nu}\overline{\mathcal{M}}^{y}_{S''}(\lambda'')). \tag{3}$$

Here ∂ represents the top dimensional stratum of the boundary, the summation is taken over all $y \in \mathrm{Crit}(f)$, all partitions of S into two subsets S', S'', all splittings of λ as sum of two classes λ', λ'' and all the ways to insert y in S' so as to get $\langle S', y\rangle$ (this is relevant if y is already present in S'; the number of these ways is $\ell+1$ where ℓ is the number of appearances of y in S', and corresponds geometrically to the choices of Morse gluings at y). We also assume here that the set S does not contain a repetition of a critical point of odd degree. This ensures that ${}^{\nu}\overline{\mathcal{M}}^{x}_{S}(\lambda)$ admits an orientation and the above formula is verified by taking into account orientations with certain signs affecting the terms on the right hand side, while the orientation on the left side is, of course, the one induced on the boundary.

By Gromov's compactness theorem, only a finite number of the moduli spaces appearing on the right hand side of the last equation are non-empty.

This equation implies that the square of the cluster differential, as defined above, vanishes so that we deduce the first part of the following result:

Theorem 2. *The map d satisfies $d^2 = 0$. The resulting homology $\mathcal{C}\ell\, H_*(L)$ does not depend, up to isomorphism, on the choices of (f, g), J and ν.*

Notice that this statement implies that $\mathcal{C}\ell\, H_*(L) \simeq \mathcal{C}\ell\, H_*(\phi(L))$ for any symplectic diffeomorphism $\phi\colon M \to M$.

Due to the commutative DGA setting, proving the invariance part of this statement is more delicate than just the usual Floer type construction for comparison morphisms. Although the comparison morphism itself between two sets of auxiliary data is defined much as in the classical way, we need a spectral sequence argument to prove that it does induce an isomorphism on homology. Since this spectral sequence plays an important role in the theory, here is its description.

Denote by ε_D the infimum of $\int_{D^2} u^*\omega$ over the set of maps $u\colon (D^2, S^1) \to (M, L)$ which are J_t-holomorphic for some t in a compact set (for instance the set of J_t's in a one-parameter family joining the two auxiliary data J and J') and non-constant. This number is strictly positive and we use it to define the *weight* of a monomial $w(x_1 \dots x_k e^{\lambda}) = k + 2\frac{\omega(\lambda)}{\varepsilon_D}$. Then consider the following *word-area* filtration of the cluster complex $\mathcal{C}\ell\,(L; J, (f, g))$:

$$F^{\ell}\mathcal{C}\ell\,(f) = \mathbb{Q}\langle m = x_1 x_2 \dots x_k e^{\lambda} \mid w(m) \geq \ell\rangle.$$

Notice that the cluster differential preserves this filtration. If $m \in \mathcal{C}\ell\,(L; J, (f, g))$ is a monomial, then we may write $dm = d_0 m + \sum_i m_i$ with d_0 the Morse differential and the m_i are monomials with $w(m_i) \geq w(m) + 1$.

The spectral sequence $E^r(f)$ associated to the filtration $F^{\ell}\mathcal{C}\ell\,(f)$ is the *word-area spectral sequence*. The total vector space of the term $E^1(f)$ is isomorphic to $(S(s^{-1}H_*(L; \mathbb{Q})) \otimes \Lambda)^{\wedge}$ because the 0-order differential in the spectral sequence coincides with the Morse differential. It is this spectral sequence that enables us to establish the invariance of the cluster homology.

Remark 3.1. Note that, if the minimal Maslov number of L is at least 2, one may define a simpler version of the cluster complex defined on $\mathbb{Q}\langle \mathrm{Crit}(f)\rangle$ instead of its graded symmetric algebra, with the same Novikov ring, in which the differential is defined by counting only *linear* cluster trees joining two critical points, i.e. strings of gradient flowlines and J-holomorphic discs starting at x and ending at y. In this case, the word-area spectral sequence boils down to Oh's spectral sequence [18] that was especially useful in recent works of Biran.

Remark 3.2. a. A critical point y of index 0 can *never* appear as end in a 0-dimensional, non-empty moduli space of type

$$^{\nu}\mathcal{M}^{x}_{\dots,y,\dots}(\lambda)$$

(except for usual Morse flow lines). This is because, if there were a negative gradient flowline from a non-constant loop in L to a local minimum, there would be a one-parameter family of these (in our construction all J-spheres that might appear are attached to some non-trivial J-disk which means that we may indeed assume the loop in question to be non-constant). Similarly, for such a moduli space, x cannot be a local maximum.

b. An element in the cluster complex, $\tau \in \mathcal{C}\ell\,(L; J, (f, g))$, is written as a sum $\tau = \sum_\lambda (a(\lambda) + m(\lambda))e^\lambda$ where $a(\lambda) \in \mathbb{Q}$ and $m(\lambda)$ is a sum of words (in the letters consisting of the critical points of f) of length at least one. We call each of the terms $a(\lambda)e^\lambda$ with $a(\lambda) \neq 0$ *a free term* of τ. If there is a critical point x of f whose differential contains at least one free term, we say that the *complex has free terms*. Further, if the Morse index of x above is larger or equal to 1, we say that *the cluster complex has high free terms*. Notice that, due to the fact that $\mu(\lambda)$ is even, if a critical point x verifies $dx = a_0 e^\lambda + \cdots$, $a_0 \neq 0$, then $\mathrm{ind}_f(x)$ is even. Moreover, in view of point a., $\mathrm{ind}_f(x) \neq \dim(L)$.

c. It is not difficult to verify that if the cluster complex is acyclic, i.e. $\mathcal{C}\ell\, H_*(L) = 0$, then the cluster has free terms. If the cluster complex $\mathcal{C}\ell\,(L; J, (f, g))$ has high free terms, then there are, moreover, some J-disks with Maslov index ≤ 0.

Example 3.3. If S^1 is a circle in $\mathbb{C}$ we have

$$\mathcal{C}\ell\, H_*(S^1) = 0.$$

Indeed, take on S^1 the perfect Morse function with one minimum m and one maximum M. There exists one pseudoholomorphic disk passing through m, of Maslov index 2, with a class in Λ that we will denote by λ_0. For the maximum, we have $dM = 0$. The differential of the minimum can be seen to be given by $dm = (1 + M + s)e^{\lambda_0}$ where s is a polynomial in M without constant or linear terms. This implies the claim because in our ring we may then find a series Q so that $Q(1 + M + s) = 1$ which means $d(Qme^{-\lambda_0}) = 1$ (and of course, the cluster homology vanishes identically iff 1 is a boundary).

Example 3.4. In the absence of bubbling (for example if $\omega|_{\pi_2(M,L)} = 0$), then

$$\mathcal{C}\ell\, H_*(L; J, (f, g)) \simeq S((s^{-1}H_*(L; \mathbb{Q})) \otimes \Lambda)^\wedge.$$

This happens because in this case the only component of the cluster differential is provided by the usual Morse differential.

4. Fine Floer homology

The purpose of this paragraph is to introduce the *fine Floer homology*, denoted $I\!F\,H_*(-)$, which was announced earlier in these notes.

4.1. Coefficient ring and moduli spaces. To define the fine Floer complex

$$I\!FC(L_0, L_1, \eta; J, (f_0, g_0), (f_1, g_1))$$

we first recall that the choices of orientations and of a relative spin structure for L_0, L_1 have been made and are included in the notation L_0, L_1. Besides this, we need auxiliary data as follows. First, as before, we need an almost complex structure J, Morse–Smale pairs (f_i, g_i) on L_i and coherent choices of perturbations. We also assume the f_i in generic position with respect to the intersection points $L_0 \cap L_1$ in the sense that these intersection points are included in the unstable manifolds of critical points of index 0 of f_i. We denote by $\Gamma = \{\alpha\colon [0, 1] \to M : \alpha(i) \in L_i, i = 0, 1\}$ the space of continuous paths from L_0 to L_1. Here, η is an element in Γ – its choice means that we fix a basepoint for this space. We denote by Γ_η the connected component of Γ which contains η. We denote by $I(L_0, L_1)$ the intersection points between L_0 and L_1 and we let I_η be those intersection points which, viewed as constant paths, belong to Γ_η. The generators of the fine Floer complex will be precisely the elements of I_η. Up to a shift in degrees, the resulting fine Floer homology will only depend on L_0, L_1, the connected component of η and the choice of orientations and relative spin structures of L_0, L_1.

To continue the construction, note that there are two group morphisms

$$\omega\colon \pi_1\Gamma_\eta \to \mathbb{R}, \quad \mu\colon \pi_1\Gamma_\eta \to \mathbb{Z},$$

the first given by integration of ω and the second, a Maslov index type morphism, obtained in the usual way as in Robbin–Salamon [20].

We now define

$$\mathcal{R} = (S\mathbb{Q}\langle s^{-1}(\mathrm{Crit}(f_0) \cup \mathrm{Crit}(f_1))\rangle \otimes \bar{\Lambda})^\wedge \tag{4}$$

where $\bar{\Lambda}$ is the rational group ring of $\Pi = \mathrm{Im}(\omega \times \mu)$; the completion is as in the case of the cluster complex except that we take into consideration both critical points of f_0 and of f_1.

Notice that there are injective group morphisms $\phi_i : \pi_2(M, L_i)/\sim \to \Pi$ which are obtained by first assuming that η joins the base points in L_0 and L_1 and then viewing a disk with boundary in, say, L_0 as a cylinder whose end on L_1 is constant. Therefore, if we denote by Λ_i the group ring of $\pi_2(M, L_i)/\sim$, we have injective ring morphisms $\phi_i : \Lambda_i \to \bar{\Lambda}$. Thus, $\mathcal{R}$ is isomorphic to the obvious completion of

$$\mathcal{C}\ell\,(L_0; J, (f_0, g_0)) \otimes \mathcal{C}\ell\,(L_1; J, (f_1, g_1)) \otimes_{\Lambda_0 \otimes \Lambda_1} \bar{\Lambda}.$$

4.2. The fine Floer complex. This is the free differential graded module over $\mathcal{R}$ given by

$$I\!FC(L_0, L_1, \eta; J, (f_0, g_0), (f_1, g_1)) = (\mathcal{R} \otimes \mathbb{Q}\langle I_\eta\rangle, d_F).$$

The grading of the elements in I_η is obtained as follows: we consider lifts $\bar{a} \in \tilde{\Gamma}_\eta$ of the points $a \in I_\eta \subset \Gamma_\eta$ where $\tilde{\Gamma}_\eta$ is the regular covering of Γ_η associated to Π (here, as usual, the last inclusion means that we view intersection points as constant paths) and we define $|a| = \mu(\bar{a})$. This grading depends on the choices of the lifts, but different choices produce isomorphic complexes.

We now describe the differential d_F. We order the critical points in $\mathrm{Crit}(f_i)$. The differential d_F verifies the Leibniz formula and for an element $a \in I_\eta$ it is of the form:

$$d_F a = \sum_{\substack{\lambda,\ b,\ k\geq 0,\ l\geq 0 \\ (x_1,\dots,x_k,y_1\dots,y_l)}} w^a_{x_1,\dots,x_k,y_1,\dots,y_l;b}(\lambda)\, x_1 \dots x_k y_1 \dots y_l b e^\lambda$$

where the x_i's belong to $\mathrm{Crit}(f_0)$, the y_j's to $\mathrm{Crit}(f_1)$, they respect the order, $\lambda \in \Pi$, and finally $b \in I_\eta$.

The coefficients $w^a_{x_1,\dots,x_k,y_1,\dots,y_l;b}(\lambda) \in \mathbb{Q}$ count the number of elements in certain 0-dimensional moduli spaces $\mathcal{W}^a_{x_1,\dots,x_k,y_1,\dots,y_l;b}(\lambda)$ (again after perturbation). These moduli spaces are defined in a way similar to the $\mathcal{M}_{\dots}(\lambda)$'s of §3. The starting point consists again of trees as in §3 but the root vertex of the tree no longer corresponds to a J-disk but rather to a J-strip (as in the usual Floer theory) which relates the two intersection points $a, b \in I_\eta$. Except for codimension two phenomena, all of the other vertices correspond to pseudoholomorphic disks with boundaries on one of the L_i's. Moreover, the gradient flows appearing in the construction correspond to one of the two functions f_i. In short, the elements of these moduli spaces are cluster trees on L_0 and L_1 that originate *at finite points* of the root strip. There may be finitely many such clusters attached to the boundary of the strip. These objects are called *cluster-strips*. It is easy to see how to associate a class $\lambda \in \Pi$ to such an object.

For generic choices of J, (f_i, g_i) and after perturbation, the dimension of the moduli space $\mathcal{W}^a_{x_1,\dots,x_k,y_1,\dots,y_l;b}(\lambda)$ is:

$$|a| - |b| - \sum |x_i| - \sum |y_j| + \mu(\lambda) - 1.$$

For one-dimensional moduli spaces, there is a formula analogous to (3). As a consequence, we have:

Theorem 3. *With the notation above, we have $d_F^2 = 0$. The resulting homology is called the fine Floer homology, $I\!F H_*(L_0, L_1, \eta)$. Up to isomorphism (and a possible shift in degrees) it does not depend on the choices made in its construction and if $\phi\colon M \to M$ is a Hamiltonian diffeomorphism, then we have isomorphisms*

$$I\!F H(L_0, L_1, \eta) \simeq I\!F H(\phi(L_0), L_1, \eta')$$

where η' corresponds to η via the Hamiltonian diffeomorphism.

Verifying $d_F^2 = 0$ is less immediate than for the cluster differential because, besides the usual breaking of clusters and of strips, there is a third potential way for

boundary points to emerge: they correspond to some cluster tree attached to a strip at some moving point p which "slides" along the boundary of the strip to one of the ends of the strip. There are two reasons that make these boundary components disappear, one is purely algebraic and is a cancellation resulting from our graded *c*ommutative setting and the other one is analytic and consists in the fact that (as remarked by Oh [16]) the usual gluing argument applies (under generic conditions) to a J-disk passing (transversally) through a and to a itself viewed as a constant strip and produces a non-constant strip with the two ends at a. In contrast with [16], our cancellation argument applies without any hypothesis of symmetry between L_0 and L_1.

4.3. Symmetrization. A particular variant of the construction of the fine Floer homology is useful in applications. To describe it, assume that L_0 equals L_1 (we denote both Lagrangian submanifolds by L), and consider a generic time-dependent Hamiltonian $H_{t\in[0,1]}$, with Hamiltonian flow $\phi_{t\in[0,1]}$, and a generic family of compatible almost complex structures $J_{t\in[0,1]}$. Take as generators of the complex the trajectories I_η^H of ϕ_t starting at time 0 and ending at time 1 on L. We may choose the generic family $J_{t\in[0,1]}$ so that $J_0 = J_1$ (we will denote this almost complex structure by J).

The *symmetric fine Floer complex* appears in this setting by additionally choosing the pair (f_0, g_0) equal to the pair (f_1, g_1) (we denote both pairs by (f, g)). Since we have a differential graded algebra multiplication map:

$$\mathcal{C}\ell\,(L; J, (f, g)) \otimes \mathcal{C}\ell\,(L; J, (f, g)) \to \mathcal{C}\ell\,(L; J, (f, g))$$

and because $J_0 = J_1 = J$, we may replace the ring

$$\mathcal{R} = (\mathcal{C}\ell\,(L_0; J_0, (f_0, g_0)) \otimes \mathcal{C}\ell\,(L_1; J_1, (f_1, g_1)) \otimes \bar{\Lambda})^\wedge$$

which appears naturally in the definition of the fine Floer complex by the ring $\hat{\mathcal{R}} = (\mathcal{C}\ell\,(L; J, (f, g)) \otimes_\Lambda \bar{\Lambda})^\wedge$. To define a differential on

$$\hat{\mathcal{R}} \otimes \mathbb{Q}\langle I_\eta^H \rangle$$

we use moduli spaces consisting of configurations in which the root pseudoholomorphic strip is replaced by a semi-cylinder satisfying the usual (J, H)-Floer type equation, with its two ends coinciding with trajectories γ, γ' and the two side boundaries lying on L.

We denote by $(\hat{I\!F}C(L; H, J, (f, g)), d_{\hat{F}})$ this symmetric fine Floer homology. If no additional notation appears the path η used in this case is just the constant path.

This homology has the same type of invariance properties as the non-symmetric version and, moreover, it is independent of the choice of $(H, J, (f, g))$ as long as it remains generic.

The advantage of this construction is that it relates directly to the cluster complex as we shall see in the following.

4.3.1. A little algebra. Consider a commutative, differential graded algebra of the form $\mathcal{A} = (SV, d)$ with V a rational vector space. Write the elements of the SV-module $SV \otimes V$ in the form

$$x_1 \dots x_k \otimes v = x_1 \dots x_k \bar{v}$$

where $v, x_i \in V$. Define the linear map

$$\alpha \colon SV \to SV \otimes V$$

by letting $\alpha(v) = \bar{v}$, $\alpha(1) = 0$ where 1 is the unit in SV and extending this map by the formula

$$\alpha(ab) = a\alpha(b) + (-1)^{|a||b|} b\alpha(a).$$

Induction on the length of words easily shows that this map is well defined and that the formula above is verified for all homogeneous monomials a, b. Explicitly, we have

$$\alpha(x_1 \dots x_k) = \sum_i (-1)^{\sigma_i} x_1 \dots \hat{x}_i \dots x_k \bar{x}_i$$

where σ_i is the product of the degree of x_i with the sum of the degrees of the x_j, $j > i$. Define the map d on $SV \otimes V$ as the unique (SV, d)-module extension of

$$d\bar{v} = \alpha(dv)$$

which verifies the standard graded Leibniz rule.

One easily checks that d so defined is a differential and that α is a chain map. Denote by

$$\widetilde{\mathcal{A}} = (SV \otimes V, d)$$

the $\mathcal{A}$-differential module constructed in this way.

4.3.2. String-strip symmetrization. We return to our geometric setting. The algebraic construction above appears in the next proposition.

Proposition 4.1. *The symmetric fine Floer homology verifies:*

$$s^{-1}\hat{I\!F} H_*(L) \simeq H_*(\widetilde{\mathcal{C}\ell}\,(L; J, (f, g))).$$

This isomorphism is proved in two steps. First, a chain complex $\mathcal{C}$ is constructed by using, instead of root semi-cylinders, (J, h)-linear clusters: they consist of pairs of critical points of the additional Morse function h related via a linear sequence of negative flow lines of h and J-disks. Of course, there are still f-clusters attached to this linear cluster. This construction is possible even when $f = h$ and in that case the resulting complex is precisely (the suspension of) $\widetilde{\mathcal{C}\ell}\,(L)$. The second step is to define a Piunikin–Salamon–Schwarz [19] map that relates the complexes $\hat{I\!F} C_*(L, H)$ and $\mathcal{C}$ and prove that it induces an isomorphism in homology.

An immediate corollary of this proposition gives the description of the symmetric fine Floer homology if no bubbling is present.

Corollary 4.2. *If* $\omega|_{\pi_2(M,L)} = 0$ *then*

$$\hat{I\!F} H_*(L) \simeq (S(s^{-1}H_*(L;\mathbb{Q})) \otimes \Lambda)^\wedge \otimes H_*(L;\mathbb{Q}).$$

Clearly, if a Lagrangian is displaceable, then both the fine Floer homology and its symmetric version vanish.

5. Applications of cluster homology

We will describe three consequences of the cluster theory developed with Cornea.

5.1. The Gromov–Sikorav problem. As a first consequence, we analyze a plausible conjecture going back to Gromov's original paper [10] on pseudoholomorphic curves, stated orally by Sikorav in the late eighties in the following way: given any compact Lagrangian submanifold of $\mathbb{C}^n$, there is a holomorphic disk passing through each point of L.

Corollary 5.1. *Let* $L \subset M$ *be a compact, orientable, relative spin Lagrangian submanifold of any symplectic manifold* M. *Assume that* $\hat{I\!F} H_*(L) = 0$ (*for example if* L *is displaceable by a Hamiltonian isotopy*). *Then, for any generic almost complex structure* J *compatible with the symplectic form, one of the following holds:*

i. *There are* J*-holomorphic disks with boundary on* L *passing through a dense subset of points of* L.

ii. *The cluster complex* $\mathcal{C}\ell\,(L, J, f)$ *has high free terms for some Morse function* f *with a single local minimum and a single local maximum* (*in particular, there are* J*-disks of non-positive Maslov index*).

Proof. Assume that for any Morse function with a single local minimum and local maximum, the associated cluster complex does not have high free terms. Fix such a function f and denote its minimum by m. By Proposition 4.1, $s^{-1}\hat{I\!F} H_*(L)$ is isomorphic to $H_*(\widetilde{\mathcal{C}\ell}\,(L, J; (f, g)))$, which means that $H_*(\widetilde{\mathcal{C}\ell}\,(L, J; (f, g)))$ vanishes. Notice that if $d\bar{m} \neq 0$, then we also have $dm \neq 0$ in the cluster complex. Assume now that

$$\bar{m} \in \widetilde{\mathcal{C}\ell}\,(L, J; (f, g))$$

is a cycle. Using Remark 3.2, we see that $\bar{m}$ can be a boundary only if $\mathcal{C}\ell\,(-)$ has free terms: indeed, by this remark, the only possible primitive of $\bar{m}$ must have the form $\tau\bar{m}$ where τ is a primitive of the unit 1 in the cluster complex. This means that there is a free term in some dx for some $x \in \mathrm{Crit}(f)$. Once again by Remark 3.2, the index of this x cannot be n and it cannot be strictly between 0 and n by our assumption. Therefore, $x = m$ and $dm \neq 0$.

The fact that dm is different from 0 means that there exists a non-empty moduli space ${}^{\nu}\mathcal{M}^m_{x_1,\ldots,x_k}(\lambda)$ of dimension 0. But for a cluster tree to originate at the minimum

m, the root disk must go through m. As we may use a different function f to place m in any generic point in L, this implies the claim. □

The dichotomy in the statement of the previous corollary can be sometimes resolved by homological restrictions.

Corollary 5.2. *Suppose that L is orientable, relative spin and that $H_{2k}(L;\mathbb{Q}) = 0$ for $2k \notin \{0, \dim(L)\}$. If $\hat{I\!F} H_*(L) = 0$, then L verifies* (i) *of Corollary* 5.1 *above.*

Example 5.3. The homological restriction in the corollary above is serious but, still, there are many examples of such manifolds: $S^1 \times S^{n-1}$ and its connected sums with itself perhaps provide the simplest examples.

Using symmetrization, we can improve these results in the displaceable case by bounding from above the area of the disks detected in terms of the displacing energy. This upper bound and Gromov's compactness theorem then imply:

Corollary 5.4. *Suppose that the relative spin, orientable Lagrangian submanifold L is displaceable by a Hamiltonian isotopy and let $E(L)$ be its Hofer displacement energy. Any ω-tame almost complex structure J has the property that one of the following is true:*

i. *For any point $x \in L$ there exists a J-holomorphic disk of symplectic area at most $E(L)$ whose boundary rests on L and which passes through x.*

ii. *There exists a J-disk of Maslov index at most*

$$2 - \min\{2k \in \mathbb{N}\backslash\{0, \dim(L)\} : H_{2k}(L;\mathbb{Q}) \neq 0\}$$

and of symplectic area at most $E(L)$.

For a Lagrangian submanifold $L \subset (M, \omega)$, define, as in Barraud–Cornea [4], its real, or relative, Gromov radius as the supremum $r(L)$ of the positive constants r so that there is a symplectic embedding of the standard ball $(B(r), \omega_0) \overset{e}{\hookrightarrow} (M, \omega)$ with the property that $e^{-1}(L) = \mathbb{R}^n \cap B(r)$. Define its real Gromov capacity $c_G(L)$ as $\pi r^2(L)/2$.

Corollary 5.5. *If an orientable, relatively spin Lagrangian with $H_{2k}(L;\mathbb{Q}) = 0$ for $2k \notin \{0, \dim(L)\}$ is displaceable, then*

$$E(L) \geq c_G(L).$$

Actually, Barraud–Cornea [4] introduced an even more relative notion of Gromov capacity: if L_0, L_1 are two Lagrangian submanifolds, one may define

$$c_G(L_0, L_1) = \pi r(L_0, L_1)^2/2$$

where $r(L_0, L_1)$ is the supremum of the radii of symplectic balls, disjoint from L_1, whose real parts rest on L_0. They show in [4] that, if L_1 is the Hamiltonian image of L_0, and assuming that we are in a case where real bubbling off does not occur, an analogous energy–capacity inequality holds: $E \geq c_G(L_0, L_1)$ where E is the Hofer energy of the pair (L_0, L_1).

5.2. Constraints on Maslov indices. The cluster complex setting provides straightforward proofs of various constraints regarding Maslov indices of Lagrangian submanifolds. See the section on applications in [5]. However, it has not yet provided a proof of the Audin conjecture, except in dimension 2. Such an application is still premature before the algebraic operations on cluster homology have been well understood.

5.3. Detection of periodic orbits. In the presence of an orientable relative spin pair of Lagrangian submanifolds L, L', we show that, by replacing in the definition of the clustered moduli spaces one (and only one) of the J-disks by a pseudoholomorphic cylinder with one boundary on L and the other boundary on L', one can construct a chain map:

$$\mathrm{cyl} : \mathcal{C}\ell\,(L) \to \mathcal{C}\ell\,(L) \otimes \mathcal{C}\ell\,(L') \otimes \Lambda_{\Phi_0}$$

where Λ_{Φ_0} is an appropriate Novikov ring.

This map induces a morphism in homology whose non-triviality is used to detect the existence of non-constant periodic orbits of Hamiltonian diffeomorphisms that *separate* L from L'. These results, obtained in [5], generalize those in [9]. When L and L' are disjoint sections in a cotangent bundle, the non-triviality is easy to detect. But, in a general symplectic manifold, such a non-triviality could be detected by making use of a specific Hamiltonian H that does have periodic orbits in the prescribed classes, and using the Albers map from the symplectic Floer (or Floer–Novikov) homology of the ambient space (using H) to each of the cluster homologies of L and L'. The gluing of the two half cylinders at each of the orbits of H would then, plausibly, provide a non-trivial realisation of the above morphism, that could then be applied to prove the existence of orbits of any other separating Hamiltonian.

5.4. Concluding remarks on cluster homology. What remains to be done on the cluster theory is: (1) to establish the analysis of its moduli spaces on a solid ground, and (2) make it computable in non-trivial examples. Although (1) has not been entirely written down at the time of submitting these notes, it seems very likely that the transversalities issues fit very well in the Hofer–Wysocki–Zehnder polyfold setting so that, hopefully, (1) will be accomplished shortly after the HWZ approach fully appears in print. A second scheme to solve similar transversality problems in the absolute case has been very recently suggested by Cieliebak–Mohnke and could be likely adapted to the cluster context thanks to the existence of a Donaldson symplectic hypersurface in the complement of any Lagrangian L (established by Auroux–Gayet–Mohsen in [3]) whose role would be to stabilize the holomorphic discs (one would then use the abstract space of non-embedded cluster trees as the new parameter space in defining the dependence of the almost complex structure J). The first step in (2) will be to establish a formula for the cluster homology of the surgery on two Lagrangian submanifolds intersecting transversally. Such a formula requires understanding how to resolve the singularities of holomorphic surfaces with corners at the intersecting points – this was first explored in an appendix of FOOO [8].

Cluster homology may also play a role in the development of the new subject of relative (or Lagrangian) symplectic field theory.

6. The emerging field of real symplectic topology

Real symplectic topology is the study of triples (M, ω, c_M) where (M, ω) is a symplectic manifold and $c_M : M \to M$ is an anti-symplectic involution. It is easy to see that the fixed point set $\mathbb{R}M$ of c_M, called the real part of M, is a Lagrangian submanifold (when it is not empty). It is shown in [24] (see also [8]) that the space $\mathbb{R}\mathcal{J}(M)$ of ω-tame almost complex structures J on M for which c_M is anti-holomorphic is as generic as one could hope: it is both generic for the study of the full space of J-holomorphic curves in M, as well as for the subspace of *real* J-curves, i.e. the space of J-curves that are c_M-invariant.

Note that the results of the preceding sections always assumed that the Lagrangian submanifolds are orientable and relative spin. In real symplectic topology, the Lagrangian submanifold $\mathbb{R}M$ is often not orientable, for instance $\mathbb{R}P^2 \subset \mathbb{C}P^2$. Moreover, the "doubling construction" that assigns a c_M-invariant J-holomorphic 2-sphere to a disk with boundary on L only uses one of the anti-symplectic involutions of the 2-sphere, namely the reflection through the equator (leaving aside the antipodal map). For these reasons, there seems to be no point in studying real symplectic topology solely from the methods defined in the preceding sections.

A very interesting and recent development, due to Welschinger [24], provides a well-defined Gromov–Witten type invariant in real 4-dimensional symplectic manifolds for the space of real rational J-curves (for a generic J in $\mathbb{R}\mathcal{J}(M)$) in a class d passing through the right number ℓ of points in order to cut its dimension down to zero, assuming that these points are invariant under c_M. Denote by δ the number of double points of such a curve. Both ℓ and δ are topological invariants depending only on d. Let $r \leq \ell$ be the number of points belonging to $\mathbb{R}M$. For such a real rational J-curve, let $0 \leq m \leq \delta$, called the *mass*, be the number of its real isolated double points (i.e. those which belong to $\mathbb{R}M$ and which are isolated in $\mathbb{R}M$, being the intersection of complex conjugated branches of the curve). Finally, let $n_d(m)$ be the number of such real curves in class d passing through that generic set of points with m real isolated double points. Then the number $\chi_r^d = \sum_{m=0}^{\delta} (-1)^m n_d(m)$ is independent of both the choice of generic $J \in \mathbb{R}\mathcal{J}(M)$ and the c_M-invariant set of points, as long as the number r remains unchanged. Formally summing over r provides an invariant of the deformation class of (M, ω, c_M). Note that, obviously, χ_r^d is then a lower bound for the number of real rational curves, in particular χ_ℓ^d is a lower bound for the number of real rational curves in class d passing through ℓ real points. These numbers have been computed in various cases, notably by Mikhalkin for toric real surfaces (see his contribution in these proceedings).

These results have been extended to higher dimensions by Welschinger in [25]. Basically, these invariants are obtained by a procedure which has some similarities

with the cluster complex, since they are extracted by gluing together various moduli spaces in an appropriate way. However, they do not point in the direction of a cluster homology, but rather suggest that there might exist some form of real quantum product related to mass. That is to say: the Lagrangian submanifolds occuring as real loci of real symplectic manifolds seem to be special enough so that one could overcome the real bubbling off phenomenon by gluing appropriate moduli spaces and extracting the desired invariants directly from these "clustered" moduli spaces much like in the complex (non-relative) case, without having to introduce a Floer or cluster complex.

References

[1] Albers, P., On the extrinsic topology of Lagrangian submanifolds. *Internat. Math. Res. Notices* **38** (2005), 2341–2371.

[2] Audin, M., Lalonde, F., Polterovich, L., Symplectic rigidity: Lagrangian submanifolds. In *Holomorphic curves in symplectic geometry* (ed. by M. Audin and J. Lafontaine), Progr. Math. 117, Birkhäuser, Basel 1994, 271–321.

[3] Auroux, D., Gayet, D., Mohsen, J.-P., Symplectic hypersurfaces in the complement of an isotropic submanifold. *Math. Ann.* **321** (2001), 739–754.

[4] Barraud, J.-F., Cornea, O., Lagrangian Intersections and the Serre spectral sequence. Preprint; arXiv:math.DG/0401094, 2004.

[5] Cornea, O., Lalonde, F., Cluster homology. Preprint; arXiv:math.SG/0508345, August 2005.

[6] Eckholm, T., Etnyre, J., Sullivan, M., Legendrian submanifolds in R^{2n+1} and contact homology. arXiv:math.SG/0210124, December 2002.

[7] Eliashberg, Y., Polterovich, L., Local Lagrangian 2-knots are trivial. *Ann. of Math.* **144** (1996), 61–76.

[8] Fukaya, K., Oh, Y.-G., Ohta, H., Ono, K., Lagrangian intersection Floer theory - anomaly and obstruction. Preprint, 2002.

[9] Gatien, D., Lalonde, F., Holomorphic cylinders with Lagrangian boundary conditions and Hamiltonian dynamics. *Duke Math. J.* **102** (2000), 485–511.

[10] Gromov, M., Pseudoholomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), 307–347.

[11] Hind, R., Lagrangian spheres in $S^2 \times S^2$. *Geom. Funct. Anal.* **14** (2004), 303–318.

[12] Hofer, H., Wysocki, K., Zehnder, E., *Polyfolds and Fredholm theory*. In preparation.

[13] Ivrii, A., Lagrangian unknottedness of tori in certain symplectic 4-manifolds. Ph.D. Thesis, Stanford, 2004.

[14] Lalonde, F., Sikorav, J.-C., Sous-variétés lagrangiennes et lagrangiennes exactes des fibrés cotangents. *Comment. Math. Helv.* **66** (1991), 18-33.

[15] McDuff, D., Salamon, D., *J-Holomorphic Curves and Symplectic Topology*. Amer. Math. Soc. Colloq. Publ. 52, Amer. Math. Soc., Providence, RI, 2004.

[16] Oh, Y.-G., Floer Cohomology of Lagrangian Intersections and Pseudo-Holomorphic Disks I. *Comm. Pure Appl. Math.* **46** (1993), 949–994.

[17] Oh, Y.-G., Relative Floer and quantum cohomology and the symplectic topology of Lagrangian submanifolds. In *Contact and symplectic geometry* (Cambridge, 1994), Publ. Newton Inst. 8, Cambridge University Press, Cambridge 1996, 201–267.

[18] Oh, Y.-G., Floer Cohomology, Spectral Sequences and the Maslov class of Lagrangian embeddings. *Internat. Math. Res. Notices* **1996** (7) (1996), 305–346.

[19] Piunikhin, S., Salamon, D., Schwarz, M., Symplectic Floer-Donaldson theory and quantum cohomology. In *Contact and symplectic geometry* (Cambridge, 1994), Publ. Newton Inst. 8, Cambridge University Press, Cambridge 1996, 171–200.

[20] Robbin, J., Salamon, D., The Maslov index for paths. *Topology* **32** (1993), 827–844.

[21] Schwarz, M., A quantum cup-length estimate for symplectic fixed points. *Invent. Math.* **133** (1998), 353–397.

[22] Viterbo, C., A new obstruction to embedding Lagrangian tori. *Invent. Math.* **100** (1990), 301–320.

[23] Viterbo, C., Exact Lagrange submanifolds, periodic orbits and the cohomology of free loop spaces. *J. Differential Geom.* **47** (1997), 420–468.

[24] Welschinger, J.-Y., Invariants of real symplectic 4-manifolds and lower bounds in real enumerative geometry. *Invent. Math.*, to appear.

[25] Welschinger, J.-Y., Enumerative invariants of strongly semipositive real symplectic manifolds. arXiv:math.AG/0509121, September 2005.

Université de Montréal, Département de mathématiques et de statistique, C.P. 6128 Succ. Centre-ville, Montréal H3C 3J7, Québec, Canada

E-mail: lalonde@dms.umontreal.ca

Gromov–Witten invariants and moduli spaces of curves

Xiaobo Liu *

Abstract. The purpose of this note is to explain how much information of Gromov–Witten invariants of compact symplectic manifolds are determined by the geometry of moduli spaces of curves. In the case when a manifold has semisimple quantum cohomology, we believe that the information obtained this way might determine all higher genus Gromov–Witten invariants and could be used to study the Virasoro conjecture of Eguchi–Hori–Xiong and S. Katz.

Mathematics Subject Classification (2000). Primary 53D45; Secondary 14N35.

Keywords. Gromov–Witten invariants, quantum cohomology, moduli space of curves, tautological ring.

1. Introduction

Let V be a compact symplectic manifold with an almost complex structure which is compatible with its symplectic structure. Gromov–Witten invariants of V are certain intersection numbers on the moduli spaces of stable pseudo-holomorphic curves in V. Such invariants do not depend on the choice of the almost complex structure and are therefore symplectic invariants. The generating functions of Gromov–Witten invariants satisfy many interesting partial differential equations. A large class of such equations come from the study of the moduli spaces of stable curves $\overline{\mathcal{M}}_{g,k}$ which were introduced by Deligne and Mumford in [7]. There are natural geometric ways to construct cohomology classes on $\overline{\mathcal{M}}_{g,k}$. These classes generate a ring called the tautological ring. There is a canonical way to produce differential equations for generating functions of Gromov–Witten invariants from relations in the tautological ring of $\overline{\mathcal{M}}_{g,k}$. Such equations hold for the Gromov–Witten theory of all compact symplectic manifolds, and are therefore called universal equations. These equations can be used to compute Gromov–Witten invariants and study other properties of Gromov–Witten theory like the Virasoro conjecture of Eguchi–Hori–Xiong and S. Katz. The Virasoro conjecture predicts that the generating functions of Gromov–Witten invariants of smooth projective varieties are annihilated by an infinite sequence of differential operators which form a half branch of the Virasoro algebra. This is a generalization of Witten's conjecture, which was proved by Kontsevich, that the generating function of intersection numbers on $\overline{\mathcal{M}}_{g,k}$ is a τ-function of the KdV hierarchy. Moreover

*Research was partially supported by NSF grant DMS-0505835.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

universal equations also hold for generating functions of intersection numbers on moduli spaces of spin curves. These equations can be used to study the generalized Witten conjecture which predicts that such generating functions are τ-functions of Gelfand–Dickey hierarchies.

On the other hand, it is a highly non-trivial problem to find explicit relations in the tautological ring of $\overline{\mathcal{M}}_{g,k}$, especially when the genus g is high. Knowledge of Gromov–Witten invariants of compact symplectic manifolds can be used to study this problem. We believe that relations in tautological rings are manifested in the properties of Gromov–Witten invariants of various compact symplectic manifolds. In a certain sense, one may think that Gromov–Witten theory of each compact symplectic manifold V gives a representation of the tautological ring of $\overline{\mathcal{M}}_{g,k}$. A good understanding of Gromov–Witten invariants of a class of manifolds might be sufficient to determine many relations in tautological rings.

In this expository article we will explain how much we know about universal equations and how much information can be obtained from such equations.

2. Tautological relations and universal equations

Let V^{2d} be a compact symplectic manifold with symplectic form ω and almost complex structure J such that the bilinear form $\omega(\cdot, J\cdot)$ defines an Riemannian metric on V. For simplicity, we assume that $H^{\text{odd}}(V;\mathbb{C}) = 0$. For each $A \in H_2(V;\mathbb{Z})$ and $g, k \in \mathbb{Z}_{\geq 0}$, let $\overline{\mathcal{M}}_{g,k}(V, A)$ be the moduli space of stable maps whose elements are of the form $(C; x_1, \ldots, x_k; f)$ where C is a genus-g complex curve with at most nodal singularities, $x_1, \ldots, x_k \in C$ are distinct smooth points which are called marked points, and f is a pseudo-holomorphic map from C to V such that $f_*[C] = A$ and f is stable in the sense that there is no infinitesimal deformation of f without moving marked points and their images. For each $i \in \{1, 2, \ldots, k\}$, let $\text{ev}_i : \overline{\mathcal{M}}_{g,k}(V, A) \longrightarrow V$ be the evaluation map defined by

$$\text{ev}_i(C; x_1, \ldots, x_k; f) := f(x_i)$$

and let $E_i \longrightarrow \overline{\mathcal{M}}_{g,k}(V, A)$ be the tautological line bundle whose geometric fiber over $(C; x_1, \ldots, x_k; f)$ is given by $T^*_{x_i}C$. For any $\gamma_1, \ldots \gamma_k \in H^*(V;\mathbb{C})$ and $n_1, \ldots, n_k \in \mathbb{Z}_{\geq 0}$, the associated Gromov–Witten invariant is defined to be

$$\langle \tau_{n_1}(\gamma_1) \ldots \tau_{n_k}(\gamma_k) \rangle_{g,A} := \int_{[\overline{\mathcal{M}}_{g,k}(V,A)]^{\text{virt}}} \bigcup_{i=1}^{k} (c_1(E_i)^{n_i} \cup \text{ev}_i^*(\gamma_i)),$$

where $[\overline{\mathcal{M}}_{g,k}(V, A)]^{\text{virt}}$ is the virtual fundamental class of degree

$$2\{(d-3)(1-g) + c_1(V)(A) + k\}$$

(cf. [30], [31], and [3]).

To define the generating functions, we need to fix a basis $\{\gamma_1, \dots, \gamma_N\}$ of $H^*(V; \mathbb{C})$ with γ_1 equal to the identity of the cohomology ring of V. For each symbol $\tau_n(\gamma_\alpha)$ we associate a parameter t_n^α. The collection of all such parameters is denoted by

$$t = (t_n^\alpha \mid n \in \mathbb{Z}_{\geq 0},\ \alpha = 1, \dots, N).$$

We can think of these parameters as coordinates on an infinite dimensional vector space called the *big phase space*. The subspace $\{t \mid t_n^\alpha = 0 \text{ if } n > 0\}$ is called the *small phase space*. Note that the small phase space can be canonically identified with $H^*(V; \mathbb{C})$. We also need the Novikov ring which is the completion of the multiplicative ring generated by monomials $q^A := d_1^{a_1} \dots d_r^{a_r}$ over the ring of rational numbers, where $\{d_1, \dots, d_r\}$ is a fixed basis of $H_2(V; \mathbb{Z})$ and $A = \sum_{i=1}^r a_i d_i$. The generating function of genus-g Gromov–Witten invariants is then defined by

$$F_g(t) := \sum_{k \geq 0} \frac{1}{k!} \sum_{\substack{\alpha_1, \dots, \alpha_k \\ n_1, \dots, n_k}} t_{n_1}^{\alpha_1} \dots t_{n_k}^{\alpha_k} \sum_{A \in H_2(V;\, Z)} q^A \langle \tau_{n_1}(\gamma_{\alpha_1}) \dots \tau_{n_k}(\gamma_{\alpha_k}) \rangle_{g, A}.$$

The function F_g is understood as a formal power series of t with values in the Novikov ring.

For $k,\ g \geq 0$, define a k-tensor $\langle\!\langle \underbrace{\cdots}_{k} \rangle\!\rangle_g$ to be the k-th covariant derivative of F_g with respect to the trivial connection on the big phase space. More precisely,

$$\langle\!\langle \mathcal{W}_1 \dots \mathcal{W}_k \rangle\!\rangle_g := \sum_{m_1, \alpha_1, \dots, m_k, \alpha_k} f^1_{m_1, \alpha_1} \dots f^k_{m_k, \alpha_k} \frac{\partial^k}{\partial t_{m_1}^{\alpha_1} \dots \partial t_{m_k}^{\alpha_k}} F_g \tag{1}$$

for vector fields $\mathcal{W}_i = \sum_{m,\alpha} f^i_{m,\alpha} \frac{\partial}{\partial t_m^\alpha}$ where $f^i_{m,\alpha}$ are functions on the big phase space. This tensor is called the *k-point* (*correlation*) *function*. We will identify $\frac{\partial}{\partial t_n^\alpha}$ with $\tau_n(\gamma_\alpha)$ and set $\tau_0(\gamma_\alpha) = \gamma_\alpha$ and $\tau_n(\gamma_\alpha) = 0$ if $n < 0$. We call γ_α a *primary vector field*, and $\tau_n(\gamma_\alpha)$ a *descendant vector field* with descendant level n. We use τ_+ and τ_- to denote the operator which shift the level of descendants, i.e.

$$\tau_\pm\Big(\sum_{n,\alpha} f_{n,\alpha} \tau_n(\gamma_\alpha)\Big) = \sum_{n,\alpha} f_{n,\alpha} \tau_{n\pm 1}(\gamma_\alpha)$$

where $f_{n,\alpha}$ are functions on the big phase space. Let

$$\eta_{\alpha\beta} = \int_V \gamma_\alpha \cup \gamma_\beta$$

be the intersection form on $H^*(V, \mathbb{C})$. We will use $\eta = (\eta_{\alpha\beta})$ and $\eta^{-1} = (\eta^{\alpha\beta})$ to lower and raise indices. For example,

$$\gamma^\alpha := \eta^{\alpha\beta} \gamma_\beta.$$

Here we are using the summation convention that repeated indices (in this formula, β) should be summed over their entire ranges.

When V is a point, the moduli space $\overline{\mathcal{M}}_{g,k}(\{\mathrm{pt}\}, 0)$ is exactly the moduli space of genus-g stable curves with k-marked points constructed by Deligne and Mumford [7]. This space is usually denoted by $\overline{\mathcal{M}}_{g,k}$. Let ψ_i be the first Chern class of the tautological line bundle E_i over $\overline{\mathcal{M}}_{g,k}$. These classes are called ψ-classes. There are natural forgetful maps $\overline{\mathcal{M}}_{g,k+n} \longrightarrow \overline{\mathcal{M}}_{g,k}$ which forgets the last n marked points of a stable curve. There are also two types of natural gluing maps between moduli spaces of stable curves. The first type is $\overline{\mathcal{M}}_{g_1,k_1+1} \times \overline{\mathcal{M}}_{g_2,k_2+1} \longrightarrow \overline{\mathcal{M}}_{g_1+g_2,k_1+k_2}$ which glues the last marked points of two stable curves to form a new stable curve. The second type is $\overline{\mathcal{M}}_{g,k+2} \longrightarrow \overline{\mathcal{M}}_{g+1,k}$ which glues the last two marked points on a stable curve to form a new stable curve. The tautological ring of $\overline{\mathcal{M}}_{g,k}$, denoted by $R^*(\overline{\mathcal{M}}_{g,k})$, is the smallest $\mathbb{Q}$-subalgebra of the Chow ring of $\overline{\mathcal{M}}_{g,k}$ which contains all ψ-classes and is closed under the push-forward of forgetting maps and gluing maps. There is also a natural stratification of $\overline{\mathcal{M}}_{g,k}$ which is labeled by dual graphs of stable curves. The dual graph G of a stable curve C is defined in the following way: The vertices of G are irreducible components of C labeled by their genera. Two vertices are connected by an edge if the corresponding irreducible components intersects at a node of C. Therefore edges of G are one to one correspondent to nodes of C. Finally each marked point of C gives a tail of G emanating from a vertex whose corresponding irreducible component of C contains this marked point. The set of all genus-g stable curves with k marked points whose dual graph is isomorphic to G is denoted by $\mathcal{M}_{g,k}(G)$. The closure of this space, i.e. $\overline{\mathcal{M}}_{g,k}(G)$, gives a class in the tautological ring $R^*(\overline{\mathcal{M}}_{g,k})$. Such classes are called boundary classes. We will call any relation among powers of ψ-classes and boundary classes a *tautological relation*.

There is a natural map

$$\mathrm{St}\colon \overline{\mathcal{M}}_{g,k}(V, A) \longrightarrow \overline{\mathcal{M}}_{g,k}$$

which forgets the map f in $(C; x_1, \ldots, x_k; f) \in \overline{\mathcal{M}}_{g,k}(V, A)$ and stabilizes

$$(C; x_1, \ldots, x_k)$$

by squeezing unstable components to points. A tautological relation on $\overline{\mathcal{M}}_{g,k}$ can be pulled back to $\overline{\mathcal{M}}_{g,k}(V, A)$ which in turn gives relations between various Gromov–Witten invariants of V. Such relations can be described by partial differential equations for $F_0, \ldots, F_g$. Equations obtained in this way hold for all compact symplectic manifolds and therefore are called *universal equations*. It is very convenient to write universal equations as equations for tensors $\langle\!\langle \cdots \rangle\!\rangle_g$.

There is a canonical way to translate tautological relations to universal equations. A boundary class $\overline{\mathcal{M}}_{g,k}(G)$ in a tautological relation corresponds to a product of correlation functions in the universal equation according to the following rules: Each

tail of G is assigned an arbitrary vector field $\mathcal{W}_i$ on the big phase space. Each edge of G is assigned a pair of primary vector fields γ_α and γ^α, one for each half edge. Each genus-h vertex of G is assigned a correlation function of the form

$$\langle\!\langle \mathcal{W}_{i_1} \dots \mathcal{W}_{i_m}\ \gamma_{\alpha_1} \dots \gamma_{\alpha_n}\ \gamma^{\beta_1} \dots \gamma^{\beta_p} \rangle\!\rangle_h$$

if the corresponding tails and half edges are connected to this vertex. Then the boundary class $\overline{\mathcal{M}}_{g,k}(G)$ is assigned the product of all correlation functions associated with all vertices of G. The coefficient of this function in the universal equation should be the coefficient of $\overline{\mathcal{M}}_{g,k}(G)$ in the tautological relation divided by the number of elements in the automorphism group of G. To interpret ψ-classes in a tautological relation, an operator T was introduced in [33] which is defined by

$$T(\mathcal{W}) := \tau_+(\mathcal{W}) - \langle\!\langle \mathcal{W}\, \gamma^\alpha \rangle\!\rangle_0 \gamma_\alpha$$

for any vector field $\mathcal{W}$. If $\psi_i^{n_i}$ is involved in a tautological relation, then in the corresponding universal equation, $\mathcal{W}_i$ should be replaced by $T^{n_i}(\mathcal{W}_i)$. The reason for this is that the cohomology class $c_1(E_i) - St^*(\psi_i)$ on $\overline{\mathcal{M}}_{g,k}(V, A)$ is represented by a cycle consisting of elements $(C; x_1, \dots, x_k; f)$ where C has a genus-0 component with only one marked point, the i-th marked point, and only one node connecting this genus-0 component to another component of C (cf. [17] and [28]).

The simplest tautological relation is $\psi_1 = 0$ on $\overline{\mathcal{M}}_{0,3}$ since $\dim \overline{\mathcal{M}}_{0,3} = 0$. This relation can be translated into the universal equation

$$\langle\!\langle T(\mathcal{W}_1)\ \mathcal{W}_2\ \mathcal{W}_3 \rangle\!\rangle_0 = 0$$

for arbitrary vector fields $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$. This equation is called the genus-0 topological recursion relation. It was observed in [44] that derivatives of this equation gives the generalized WDVV equation:

$$\langle\!\langle \mathcal{W}_1\ \mathcal{W}_2\ \gamma^\alpha \rangle\!\rangle_0 \langle\!\langle \gamma_\alpha\ \mathcal{W}_3\ \mathcal{W}_4 \rangle\!\rangle_0 = \langle\!\langle \mathcal{W}_1\ \mathcal{W}_3\ \gamma^\alpha \rangle\!\rangle_0 \langle\!\langle \gamma_\alpha\ \mathcal{W}_2\ \mathcal{W}_4 \rangle\!\rangle_0$$

for arbitrary vector fields $\mathcal{W}_1, \dots, \mathcal{W}_4$. Because of this equation, we can define an associative product

$$\mathcal{W}_1 \circ \mathcal{W}_2 := \langle\!\langle \mathcal{W}_1\ \mathcal{W}_2\ \gamma^\alpha \rangle\!\rangle_0 \gamma_\alpha \tag{2}$$

for any vector fields $\mathcal{W}_1$ and $\mathcal{W}_2$ on the big phase space. This product is called the *quantum product on the big phase space* (cf. [33]). When restricted to the tangent bundle of the small phase space, this product is exactly the product for the quantum cohomology of V. Unlike the quantum product on the small phase space which has an identity element γ_1, the quantum product on the big phase space does not have an identity. Let

$$\mathcal{S} := -\sum_{m,\alpha} \tilde{t}_m^\alpha \tau_{m-1}(\gamma_\alpha)$$

be the *string vector field*, where $\tilde{t}_m^\alpha := t_m^\alpha - \delta_{m,1}\delta_{\alpha,1}$. The *string equation* has the form

$$\langle\!\langle \mathcal{S} \rangle\!\rangle_g = \frac{1}{2}\delta_{g,0}\eta_{\alpha\beta}t_0^\alpha t_0^\beta$$

for $g \geq 0$. By the second derivative of the genus-0 string equation, we can rewrite the operator T as

$$T(\mathcal{W}) = \tau_+(\mathcal{W}) - \mathcal{S} \circ \tau_+(\mathcal{W}).$$

Therefore T measures the difference between $\mathcal{S}$ and the identity of the quantum product on the big phase space, which actually does not exist by our definition of the quantum product. This gives an algebraic interpretation for the operator T.

Universal equations are very powerful for a class of manifolds for which the quantum product defined by equation (2) is semisimple in the following sense: Define the *Euler vector field* on the big phase space by

$$\mathcal{X} := -\sum_{m,\alpha}(m + b_\alpha - b_1 - 1)\,\tilde{t}_m^\alpha\,\tau_m(\gamma_\alpha) - \sum_{m,\alpha,\beta} C_\alpha^\beta \tilde{t}_m^\alpha\,\tau_{m-1}(\gamma_\beta),$$

where (C_α^β) is the matrix of multiplication by $c_1(V)$ in the ordinary cohomology ring of V with respect to the basis $\{\gamma_1, \ldots \gamma_N\}$ and $b_\alpha = \frac{1}{2}(\dim(\gamma_\alpha) - d + 1)$ for a compact symplectic manifold V of dimension $2d$. In case that V is a smooth projective variety, we can choose $\gamma_\alpha \in H^{p_\alpha,q_\alpha}(V)$ and b_α should be modified as $b_\alpha = p_\alpha - \frac{1}{2}(d-1)$. This modification is necessary for formulating the Virasoro conjecture below. The Euler vector field satisfies the following *quasi-homogeneity equation*

$$\langle\!\langle \mathcal{X} \rangle\!\rangle_g = (3-d)(1-g)F_g + \frac{1}{2}\delta_{g,0}\sum_{\alpha,\beta} C_{\alpha\beta}t_0^\alpha t_0^\beta - \frac{1}{24}\delta_{g,1}\int_V c_1(V) \cup c_{d-1}(V).$$

The quantum multiplication by $\mathcal{X}$ is an endomorphism on the vector space spanned by primary vector fields on the big phase space. We say that V has *semisimple* quantum cohomology if this endomorphism has distinct eigenvalues at generic points. Since our definition of quantum product on the big phase space coincides with usual quantum product when restricted to the small phase space, and the restriction of $\mathcal{X}$ to the small phase space also coincides with the Euler vector field for usual quantum cohomology, this definition of semisimplicity is a generalization of the semisimple condition used by Dubrovin in [10]. Under the semisimplicity assumption, there exists vector fields $\mathcal{E}_1, \ldots, \mathcal{E}_N$ on the big phase space which are linear combinations of primary vector fields such that

$$\mathcal{X} \circ \mathcal{E}_i = u_i\mathcal{E}_i, \quad \mathcal{E}_i \circ \mathcal{E}_j = \delta_{ij}\mathcal{E}_i$$

for every i and j where u_i are functions on the big phase space (cf. [36]). These vector fields are called *idempotents* of the quantum product on the big phase space. When restricted to the small phase space, $(u_1, \ldots, u_N)$ is precisely the *canonical coordinate system* for the semisimple Frobenius structure studied in [10]. On the big

phase space these functions are not sufficient to give a nice coordinate system. The idempotents will play the role of the canonical coordinate system. We refer to [36] for more properties of the idempotents on the big phase space.

It is well known that the quantum cohomology on the small phase space defines a Frobenius manifold structure (cf. [10]). However the quantum product on the big phase space defined be equation (2) does not give an infinite dimensional Frobenius manifold structure. To obtain a Frobenius structure on the big phase space, we should modify the definition of the quantum product in the following way: For any vector fields $\mathcal{W}_1$ and $\mathcal{W}_2$ on the big phase space define

$$\mathcal{W} \diamond \mathcal{V} := \sum_{k=0}^{\infty} \langle\!\langle \tau_-^k(\mathcal{W})\ \tau_-^k(\mathcal{V})\ \gamma^\alpha \rangle\!\rangle_0\ T^k(\gamma_\alpha).$$

This product is commutative and associative. The associativity follows from the fact that $\langle\!\langle \{\tau_-^k T^l(\gamma_\alpha)\}\ \mathcal{W}\ \mathcal{V} \rangle\!\rangle_0 = 0$ if $k \neq l$. Moreover for any primary vector fields $\mathcal{W}$ and $\mathcal{V}$,

$$T^k(\mathcal{W}) \diamond T^l(\mathcal{V}) = \delta_{kl}\ T^k(\mathcal{W} \circ \mathcal{V}).$$

This product has an identity

$$\hat{\mathcal{I}} := \sum_{k=0}^{\infty} \langle\!\langle \mathcal{I}\, \mathcal{I} \gamma^\alpha \rangle\!\rangle_0 T^k(\gamma_\alpha) = \sum_{k=0}^{\infty} T^k(\mathcal{I} \circ \mathcal{I}).$$

Define inner product on the big phase space by

$$(\mathcal{W},\ \mathcal{V}) := \sum_{k=0}^{\infty} \langle\!\langle \mathcal{I}\ \tau_-^k(\mathcal{W})\ \tau_-^k(\mathcal{V}) \rangle\!\rangle_0.$$

This is a symmetric non-degenerate bilinear form on the big phase space. Moreover,

$$(T^m(\gamma_\alpha),\ T^n(\gamma_\beta)) = \delta_{mn}\eta_{\alpha\beta}$$

for any $m, n \in \mathbb{Z}_{\geq 0}$ and $1 \leq \alpha, \beta \leq N$. The product "$\diamond$" is compatible with this inner product in the sense that

$$(\mathcal{W}_1 \diamond \mathcal{W}_2,\ \mathcal{W}_3) = (\mathcal{W}_1,\ \mathcal{W}_2 \diamond \mathcal{W}_3)$$

for all vector fields $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ and therefore define a Frobenius algebra structure on tangent spaces of the big phase space. Note that if the quantum product "$\circ$" is semisimple in the above sense and $\mathcal{E}_1, \ldots, \mathcal{E}_N$ are the idempotents of "$\circ$", then vector fields

$$\{T^n(\mathcal{E}_i) \mid n \in \mathbb{Z}_{\geq 0},\ i = 1, \ldots, N\}$$

are idempotents for "$\diamond$" and give a frame for the tangent bundle of the big phase space. This also justifies the notion of semisimplicity introduced above. We also note that

this frame is not commutative with respect to the Lie bracket and therefore does not come from coordinate vector fields for any coordinate system on the big phase space. Since the product "$\circ$" is much easier to use than the product "$\diamond$", we will not use the product "$\diamond$" in this paper.

In [27] and [42], the WDVV equation has been used to compute genus-0 Gromov–Witten invariants. Higher genus analogues of genus-0 universal equations can also be used to compute higher genus Gromov–Witten invariants. Lets first see what happens for genus-1 and genus-2 cases.

The genus-1 analogue of the genus-0 topological recursion relation is the following

$$\langle\!\langle T(\mathcal{W})\rangle\!\rangle_1 = A_0(\mathcal{W}) := \frac{1}{24}\langle\!\langle\, \mathcal{W}\, \gamma^\alpha\, \gamma_\alpha \rangle\!\rangle_0.$$

This equation is translated from the following tautological relation on $\overline{\mathcal{M}}_{1,1}$ (cf. [8]):

$$\psi_1 = \frac{1}{12}\{\text{boundary stratum of } \overline{\mathcal{M}}_{1,1}\}.$$

The genus-1 analogue of the WDVV equation is the following equation discovered by Getzler [16]: For any vector fields $\mathcal{W}_1, \ldots, \mathcal{W}_4$,

$$\begin{aligned}\sum_{\sigma\in S_4} \Big\{ & 4\langle\!\langle \{\mathcal{W}_{\sigma(1)} \circ \mathcal{W}_{\sigma(2)} \circ \mathcal{W}_{\sigma(3)}\}\, \mathcal{W}_{\sigma(4)} \rangle\!\rangle_1 \\ & - 3\langle\!\langle \{\mathcal{W}_{\sigma(1)} \circ \mathcal{W}_{\sigma(2)}\}\, \{\mathcal{W}_{\sigma(3)} \circ \mathcal{W}_{\sigma(4)}\} \rangle\!\rangle_1 \\ & + \langle\!\langle \{\mathcal{W}_{\sigma(1)} \circ \mathcal{W}_{\sigma(2)}\}\, \mathcal{W}_{\sigma(3)}\, \mathcal{W}_{\sigma(4)}\, \gamma^\alpha \rangle\!\rangle_0 \langle\!\langle \gamma_\alpha \rangle\!\rangle_1 \\ & -2\langle\!\langle \mathcal{W}_{\sigma(1)}\, \mathcal{W}_{\sigma(2)}\, \mathcal{W}_{\sigma(3)}\, \gamma^\alpha \rangle\!\rangle_0 \langle\!\langle \{\gamma_\alpha \circ \mathcal{W}_{\sigma(4)}\} \rangle\!\rangle_1 \Big\} = B_0(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3, \mathcal{W}_4)\end{aligned}$$

where B_0 is a symmetric 4-tensor which consists of 3 terms involving only genus-0 data (see, for example, [36] for the precise form of B_0 where we used notation G_0 for this tensor). This equation corresponds to a tautological relation on $\overline{\mathcal{M}}_{1,4}$.

Let F_1^s be the restriction of F_1 to the small phase space. Using the genus-1 topological recursion relation, Dijkgraaf and Witten [9] obtained a formula, called the *genus*-1 *constitutive relation*, for computing F_1 from F_1^s. In [11], Dubrovin and Zhang observed that if the quantum cohomology of V is semisimple, then in the canonical coordinate system $(u_1, \ldots, u_N)$ on the small phase space, Getzler's equation gives all second order partial derivatives $\frac{\partial^2 F_1^s}{\partial u_i \partial u_j}$ in terms of genus-0 data. They then use the quasi-homogeneity equation to obtain a formula for all first order derivatives of F_1^s. This formula determines F_1^s in terms of genus-0 functions up to an additive constant. It was observed in [36] that one can solve Getzler's equation on the big phase using idempotents $\mathcal{E}_1, \ldots, \mathcal{E}_N$ and obtain

$$\langle\!\langle\, \mathcal{E}_i \rangle\!\rangle_1 = \frac{1}{24}\Big\{ \langle\!\langle\, \tau_-(\mathcal{L}_0)\, \gamma_\alpha\, \gamma^\alpha\, \mathcal{E}_i \rangle\!\rangle_0 - \sum_j u_j\, B_0(\mathcal{E}_j, \mathcal{E}_j, \mathcal{E}_j, \mathcal{E}_i) \Big\}$$

where $\mathcal{L}_0 := -\mathcal{X} - (b_1 + 1)T(\mathcal{S})$. Let ∇ be the trivial connection on the big phase space which is uniquely characterized by the requirement that all vector fields $\tau_n(\gamma_\alpha)$ are parallel. Then $\nabla_{\mathcal{E}_i}\tau_-(\mathcal{L}_0) = 0$. Moreover, second order derivatives of the string equation imply that $A_0(\mathcal{S})$ is a constant. So the above equation can be written in the following form:

Theorem 2.1 ([36]). *For manifolds with semisimple quantum cohomology,*

$$\mathcal{E}_i F_1 = \frac{1}{24}\Big\{\mathcal{E}_i\ A_0\,(\tau_-(\mathcal{L}_0) + \mathcal{S}) - \sum_j u_j\, B_0(\mathcal{E}_j, \mathcal{E}_j, \mathcal{E}_j, \mathcal{E}_i)\Big\}$$

for $i = 1, \ldots, N$.

Together with the genus-1 topological recursion relation, this equation gives all the first order derivatives of F_1 and therefore completely determines F_1 up to a constant. We also note that when restricted to the small phase space, this equation is equivalent to the equation obtained in [11].

The genus-2 analogue of the topological recursion relation is

$$\langle\!\langle T^2(\mathcal{W})\rangle\!\rangle_2 = A_1(\mathcal{W})$$

where A_1 is a 1-tensor which consists of 5 terms involving only genus-0 and genus-1 data (see, for example, [33] for the precise form of A_1). We call this equation Mumford equation since it corresponds to a tautological relation on $\overline{\mathcal{M}}_{2,1}$ of the form

$$\psi_1^2 = \text{linear combinations of boundary strata of } \overline{\mathcal{M}}_{2,1},$$

which was proved in [40]. Mumford's tautological relation was first translated to a universal equation by Getzler [17] following some observations by Faber.

The genus-2 analogue of the WDVV equation is the following equation due to Belorousski and Pandharipande [4]: For arbitrary vector fields $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$ on the big phase space we have

$$\begin{aligned}
&2\langle\!\langle\{\mathcal{W}_1 \circ \mathcal{W}_2 \circ \mathcal{W}_3\}\rangle\!\rangle_2 - 2\langle\!\langle \mathcal{W}_1\, \mathcal{W}_2\, \mathcal{W}_3\, \gamma^\alpha\rangle\!\rangle_0 \langle\!\langle T(\gamma_\alpha)\rangle\!\rangle_2 \\
&\qquad + \frac{1}{2}\sum_{\sigma\in S_3} \langle\!\langle \mathcal{W}_{\sigma(1)}\, T(\mathcal{W}_{\sigma(2)} \circ \mathcal{W}_{\sigma(3)})\rangle\!\rangle_2 - \langle\!\langle T(\mathcal{W}_{\sigma(1)})\,\{\mathcal{W}_{\sigma(2)} \circ \mathcal{W}_{\sigma(3)}\}\rangle\!\rangle_2 \\
&\quad = B_1(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3),
\end{aligned}$$

where B_1 is a symmetric 3-tensor which consists of 16 terms involving only genus-0 and genus-1 data (see, for example, [33] for the precise form of B_1 where we used the notation B for this tensor). This equation corresponds to a tautological relation on $\overline{\mathcal{M}}_{2,3}$. There is also another genus-2 equation due to Getzler [17] which corresponds to a tautological relation on $\overline{\mathcal{M}}_{2,2}$. It was proved in [33] and [35] that Getzler's genus-2 equation can be derived from Mumford equation and Belorousski-Pandharipande equation. So we will not give the precise form of this equation. Note

that all genus-2 universal equations involve gravitational descendants and can not be reduced to equations on the small phase space. This is the main reason for introducing the quantum product on the big phase space. In [37], we proved that in the semisimple case F_2 can be solved from the above genus-2 universal equations and obtained

Theorem 2.2 ([37]). *For manifolds with semisimple quantum cohomology,*

$$F_2 = \frac{1}{6}\Big\{ A_1\left(2\,\tau_-^2(\mathcal{L}_0) + 3\,\tau_-(\mathcal{S})\right) - \sum_{i=1}^{N} u_i B_1(\mathcal{E}_i, \mathcal{E}_i, \mathcal{E}_i)\Big\}.$$

This formula is very similar to the solution of genus-1 equations if we ignore the vector field $\mathcal{E}_i$ in the formula in Theorem 2.1.

For genus bigger than 2, we do not have enough information about universal equations due to the lack of understanding of the tautological rings of moduli spaces of stable curves. So far the only non-trivial universal equation with genus bigger than 2 is the following genus-3 analogue of the topological recursion relation proved by Kimura and the author in [25]: For any vector field $\mathcal{W}$ on the big phase space

$$\langle\!\langle T^3(\mathcal{W})\rangle\!\rangle_3 = A_2(\mathcal{W})$$

where A_2 is a tensor which consists of 29 terms involving only data of genus less than or equal to 2. (See [25] for the precise form of this tensor. An equivalent formula was derived in [2] using the invariance conjecture which has not yet been proved.) This equation corresponds to a tautological relation on $\overline{\mathcal{M}}_{3,1}$ of the form

$$\psi_1^3 = \text{linear combinations of boundary strata of } \overline{\mathcal{M}}_{3,1}.$$

There is a conjecture by Getzler [17] that any degree g monomial of ψ-classes on $\overline{\mathcal{M}}_{g,k}$ should be supported on the boundary of $\overline{\mathcal{M}}_{g,k}$. This conjecture was proved by Ionel [22]. Faber and Pandharipande [15] further proved that degree g monomials of ψ-classes on $\overline{\mathcal{M}}_{g,k}$ are equal to some tautological classes on the boundary of $\overline{\mathcal{M}}_{g,k}$. Note that tautological rings also contain push-forward classes of powers of ψ-classes, i.e. the so called κ-classes, which do not appear in the definition of Gromov–Witten invariants. These results do not yet guarantee the existence of corresponding universal equations. Nevertheless, as mentioned in [33], we do expect that the following conjecture should be true.

Conjecture 2.3. For all $g \geq 1$, there is a universal equation of the form

$$\langle\!\langle T^g(\mathcal{W})\rangle\!\rangle_g = A_{g-1}(\mathcal{W})$$

where A_{g-1} is a tensor which only involve data of genus less than or equal to $g-1$.

Such equations should be the genus-g analogue of the topological recursion relation. Based on our experience from genus-1 and genus-2, we also expect that there

should be a genus-g analogue of the WDVV equation whose lower genus part should be a symmetric tensor, written as B_{g-1}, which only involves data of genus less than or equal to $g-1$. However, at this stage, we even do not have a good prediction for the top genus part of this equation for $g \geq 3$. Despite of these difficulties we believe that the following conjecture should be true.

Conjecture 2.4. For manifolds with semisimple quantum cohomology, F_g can be solved explicitly from universal equations for $g \geq 2$.

Based on Theorems 2.1 and 2.2, we may even speculate that the form of F_g obtained by solving universal equations should be

$$F_g = A_{g-1}\big(a_g\, \tau_-^g(\mathcal{L}_0) + b_g\, \tau_-^{g-1}(\mathcal{S})\big) - c_g \sum_{i=1}^{N} u_i\, B_{g-1}(\mathcal{E}_i, \ldots, \mathcal{E}_i)$$

where a_g, b_g, c_g are some constants depending only on g.

One might also expect that there exist universal equations of the form

$$\langle\!\langle T^{n_1}(\mathcal{W}_1) \ldots T^{n_k}(\mathcal{W}_k)\rangle\!\rangle_g = \text{an expression involving at most genus-}(g-1)\text{ data} \quad (3)$$

for $n_1 + \cdots + n_k = g$. This statement is somewhat stronger than Getzler's conjecture on the tautological ring of $\overline{\mathcal{M}}_{g,k}$. Since there are only finitely many strata on $\overline{\mathcal{M}}_{g,k}$ with a fixed degree, one can explicitly write out the right-hand side of equation (3) with certain undetermined coefficients. It might be possible that in many cases these coefficients can be fixed by known Gromov–Witten theory. For example, to obtain the genus-3 topological recursion relation in [25], we only need the Gromov–Witten theory of a point and $\mathbb{CP}^1$. However, coefficients obtained in this way are quite mysterious. It would be very interesting to give a better explanation to such coefficients.

3. The Virasoro conjecture

In [14] Eguchi, Hori and Xiong constructed a sequence of differential operators on the big phase space. With a slight modification proposed by S. Katz (cf. [6]), these operators satisfy the Virasoro bracket relation when the underlying manifold is a smooth projective variety and therefore form a half branch of the Virasoro algebra. They conjectured that these operators annihilate the generating function of the Gromov–Witten invariants for all smooth projective varieties. This conjecture is now known as the *Virasoro conjecture*. It is a far-reaching generalization of a conjecture of Witten, which was proved by Kontsevich, that the generating function of intersection numbers of ψ-classes on $\overline{\mathcal{M}}_{g,k}$ is a τ-function of the KdV hierarchy (cf. [45], [26], and [45]). In [13], Dubrovin and Zhang proved that if the quantum cohomology of a projective variety V is semisimple, then higher genus Gromov–Witten invariants of V are determined by genus-0 invariants and the Virasoro conjecture. For manifolds with

semisimple quantum cohomology, Givental [19] has conjectured a form of higher genus generating functions in terms of genus-0 data and the τ-function in Witten's conjecture and showed that his conjectural formula satisfies the Virasoro constraints (cf. [20]). Consequently, Givental's conjecture is equivalent to the Virasoro conjecture for projective varieties with semisimple quantum cohomology. In the case that the underlying manifold has a torus action with isolated fixed points and also has semisimple quantum cohomology, Givental has a scheme to reduce his conjecture to the so-called *R-conjecture* on the fundamental solutions to the flat section equations of a one-parameter family of connections on the small phase space defined using quantum product. Note that localization techniques played a crucial role in this scheme. An outline for the proof of the R-conjecture for projective spaces were given in [20]. The R-conjecture has been verified for flag manifolds in [24] and for Grassmannians in [5]. Also using localization techniques, Okounkov and Pandharipande [41] proved the Virasoro conjecture for algebraic curves. In this paper we will focus on the approach to the Virasoro conjecture using universal equations instead of localization techniques. Since universal equations hold for all compact symplectic manifolds, this approach should apply to a larger class of manifolds. In particular, there is no need to assume the existence of torus actions on the manifolds. In [39], Tian and the author proved the following theorem using genus-0 topological recursion relation (see also [12], [18], [34], [21] for alternative proofs.):

Theorem 3.1 ([39]). *The genus-*0 *Virasoro conjecture holds for all compact symplectic manifolds.*

The genus-1 Virasoro conjecture for manifolds with semisimple quantum cohomology was proved by Dubrovin and Zhang [12] (see also [32] and [36]). The genus-2 analogue of this result was proved in [37] using Theorem 2.2.

Theorem 3.2 ([37]). *The genus-*2 *Virasoro conjecture holds for manifolds with semisimple quantum cohomology.*

This theorem implies in particular that Givental's conjectural formula is correct in the genus-2 case. An alternative approach to the genus-2 Virasoro conjecture for manifolds with semisimple quantum cohomology is to show that Givental's formula satisfies known genus-2 universal equations and then use the result in [33] that known genus-2 universal equations uniquely determine F_2 in the semisimple case. Givental's formula was constructed using actions of twisted loop groups on a product of copies of the τ-function in Witten's conjecture. The invariance of the genus-2 Mumford's relation under the action of Lie algebras of twisted loop groups was discussed in [29]. It was also claimed that the genus-2 equations of Getzler and Belorousski-Pandharipande are invariant in the same sense. In [32] and [33], the author proved that the genus-1 and genus-2 Virasoro conjecture for all smooth projective varieties can be reduced to an SL(2) symmetry for Gromov–Witten invariants. We will give the precise statement of this result later. Below we explain in more detail the relations between universal equations and the Virasoro conjecture.

The following operators were introduced in [33] as a convenient tool in the study of the Virasoro conjecture: For any vector field $\mathcal{W} = \sum_{m,\alpha} f_{m,\alpha}\tau_m(\gamma_\alpha)$ on the big phase space, define

$$G(\mathcal{W}) := \sum_{m,\alpha,\beta} (m + b_\alpha) f_{m,\alpha}\tau_m(\gamma_\alpha), \quad C(\mathcal{W}) := \sum_{m,\alpha,\beta} f_{m,\alpha} C_\alpha^\beta \tau_m(\gamma_\beta),$$

and

$$R(\mathcal{W}) := (GT + C)(\mathcal{W}).$$

Starting from the string vector field $\mathcal{S}$, we can apply the operator R recursively to obtain a sequence of vector fields on the big phase space:

$$\mathcal{L}_n := -R^{n+1}(S)$$

for $n \geq -1$. It was proved in [33] that this sequence of vector fields satisfy the Virasoro bracket relation:

$$[\mathcal{L}_m, \mathcal{L}_n] = (m - n)\mathcal{L}_{m+n}.$$

These vector fields are not exactly the Virasoro operators given in [14] which are second order differential operators. However the Virasoro conjecture can be rephrased using these vector fields in the following way. First, second order derivatives of the genus-0 Virasoro conjecture can be reinterpreted as

$$\mathcal{L}_n \circ \mathcal{W} = -\mathcal{X}^{n+1} \circ \mathcal{W} \tag{4}$$

for any vector field $\mathcal{W}$ and $n \geq 0$, where $\mathcal{W}^k$ is defined to be the quantum product of k copies of $\mathcal{W}$. This equation follows from the associativity of the quantum product and the following property of the operator R (cf. [33]):

$$R(\mathcal{V}) \circ \mathcal{W} = \mathcal{X} \circ \mathcal{V} \circ \mathcal{W}$$

for any vector fields $\mathcal{W}$ and $\mathcal{V}$. Using this property a proof for genus-0 Virasoro conjecture was obtained in [34] which is much simpler than the original proof in [39].

To interpret higher genus Virasoro conjecture, we also need the following operators:

$$Q_0 := \tau_-, \quad Q_1 := Q := G + C\tau_-, \quad Q_k := Q(Q-1)\dots(Q-k+1)\tau_+^{k-1}$$

for $k \geq 1$. These operators were used in [34] to simplify the proof of the genus-0 Virasoro conjecture. Define second order differential operators

$$W_n := \sum_{i=1}^{n} \{Q_i(\gamma^\alpha)\} \{Q_{n-i}\tau_+ Q(\gamma_\alpha)\}$$

for $n \geq 1$. Note that $Q_i(\gamma^\alpha)$ and $Q_{n-i}\tau_+ Q(\gamma_\alpha)$ are parallel vector fields with respect to the trivial connection ∇ on the big phase space. Therefore they commute with each

other as first order differential operators. Then *Virasoro conjecture* for $g \geq 1$ can be formulated as

$$\left\{\mathcal{L}_n + \frac{1}{2}\lambda^2 W_n + \frac{1}{2}(W_n F_0)\right\} e^{\sum_{g=1}^{\infty} \lambda^{2g-2} F_g} = 0$$

for $n \geq -1$. Here $(W_n F_0)$ means the multiplication by the genus-0 function $W_n F_0$. The Virasoro conjecture for $g \geq 2$ can also be formulated as

$$\left\{\mathcal{L}_n + \frac{1}{2}\lambda^2 e^{-F_1}\, W_n\, e^{F_1}\right\} e^{\sum_{g=2}^{\infty} \lambda^{2g-2} F_g} = 0$$

for $n \geq -1$. Here we understand $W_n = 0$ for $n \leq 0$. In this formulation, $e^{-F_1}\, W_n\, e^{F_1}$ is understood as the composition of three operators acting on the space of functions.

The genus-g L_{-1}-constraint is simply the string equation which holds for all compact symplectic manifolds

$$\langle\!\langle \mathcal{L}_{-1} \rangle\!\rangle_g = -\frac{1}{2}\delta_{g,0}\eta_{\alpha\beta} t_0^\alpha t_0^\beta.$$

The genus-g L_0-constraint has the following form

$$\langle\!\langle \mathcal{L}_0 \rangle\!\rangle_g = -\frac{1}{2}\delta_{g,0} C_{\alpha\beta} t_0^\alpha t_0^\beta + \frac{1}{24}\delta_{g,1}\left\{\int_V c_1(V) \cup c_{d-1}(V) - \frac{3-d}{2}\chi(V)\right\}.$$

This constraint follows from the quasi-homogeneity equation and the *dilaton equation*

$$\langle\!\langle T(\mathcal{S}) \rangle\!\rangle_g = (2g-2)F_g + \frac{1}{24}\chi(V)\delta_{g,1}$$

where $\chi(V)$ is the Euler characteristic number of V. These equations also hold for all compact symplectic manifolds.

For $g \geq 1$ and $n \geq 1$, the genus-g L_n-constraint in the Virasoro conjecture can be formulated as:

$$\begin{aligned}\langle\!\langle \mathcal{L}_n \rangle\!\rangle_g = -\frac{1}{2}\sum_{i=1}^{n}\Big\{&\langle\!\langle Q_i(\gamma^\alpha)\ \{Q_{n-i}\tau_+ Q(\gamma_\alpha)\}\rangle\!\rangle_{g-1} \\ &+ \sum_{h=1}^{g-1} \langle\!\langle Q_i(\gamma^\alpha)\rangle\!\rangle_h \langle\!\langle \{Q_{n-i}\tau_+ Q(\gamma_\alpha)\}\rangle\!\rangle_{g-h}\Big\}.\end{aligned}$$

Note that the right-hand side of this formula only involves data of genus less than g.

For any compact symplectic manifold V, we can use universal equations to obtain recursion relations among $\langle\!\langle \mathcal{L}_n \rangle\!\rangle_g$ for $g = 1, 2$. In genus-1 case, we have (cf. [32])

$$\langle\!\langle \mathcal{L}_n \rangle\!\rangle_1 + \frac{n+1}{2}(\mathcal{S} \circ \mathcal{L}_{n-1})\langle\!\langle \mathcal{L}_1 \rangle\!\rangle_1 \equiv 0 \mod \{\text{genus 0 data}\}.$$

In genus-2 case, we have (cf. [33])

$$\langle\!\langle \mathcal{L}_n \rangle\!\rangle_2 + \frac{n+1}{2(n-1)} T(\mathcal{S} \circ \mathcal{L}_0) \langle\!\langle \mathcal{L}_{n-1} \rangle\!\rangle_2 \equiv 0 \mod \{\text{genus} \leq 1 \text{ data}\}.$$

The right-hand sides of these two equations are explicit functions only involving lower genus data (See [32] and [33] for the precise forms). As a consequence of these two formulas, we have the next result.

Theorem 3.3 ([32], [33]). *For any compact symplectic manifold, $\langle\!\langle \mathcal{L}_n \rangle\!\rangle_g$ can be computed from $\langle\!\langle \mathcal{L}_1 \rangle\!\rangle_g$ and lower genus data for all $n \geq 2$ and $g = 1, 2$.*

In case that the manifold is a projective variety we can further verify the following:

Theorem 3.4 ([32], [33]). *For any smooth projective variety, the genus-*1 *and genus-*2 *Virasoro conjecture follows from the L_1-constraint.*

Note that the first 3 Virasoro operators (as well as vector fields $\mathcal{L}_{-1}$, $\mathcal{L}_0$ and $\mathcal{L}_1$) form a 3-dimensional subalgebra which is isomorphic to SL(2). Since the L_{-1} and L_0 constraints are satisfied for all compact symplectic manifolds, the proof of the above theorem can be interpreted as saying that for genus-1 and genus-2 cases universal equations upgrade an SL(2) symmetry for Gromov–Witten invariants to the Virasoro conjecture. We believe that this should also be true for higher genus cases:

Conjecture 3.5. For all compact symplectic manifolds V, $\langle\!\langle \mathcal{L}_n \rangle\!\rangle_g$ can be computed from $\langle\!\langle \mathcal{L}_1 \rangle\!\rangle_g$ and lower genus data for $n \geq 2$ and $g \geq 1$. In case that V is a smooth projective variety the Virasoro conjecture follows from universal equations and the L_1- constraint.

Note that the above two theorems do not need semisimplicity. In case that the quantum cohomology is semisimple, we can actually recover the action of $\mathcal{L}_1$ on F_g from the action of $\mathcal{L}_n$ on F_g for $n \geq 2$ due to algebraic relations among these vector fields (cf. [32] and [33] for details). This is the reason that the genus-1 and genus-2 Virasoro conjecture hold for manifolds with semisimple quantum cohomology. In fact, one can actually use the formulas in Theorems 2.1 and 2.2 to prove the Virasoro conjecture in these cases (cf. [36] and [37] for details). Based on these results, it is reasonable to believe that the following conjecture should be true.

Conjecture 3.6. For all smooth projective varieties with semisimple quantum cohomology, the Virasoro conjecture follows from universal equations.

We also note that besides vector fields $\mathcal{L}_n$, $n \geq -1$, there are also other natural vector fields on the big phase space which satisfy the Virasoro bracket relation. For example, if we define $\overline{\mathcal{X}}^k := \mathcal{S} \circ \mathcal{X}^k$, then these vector fields satisfy the bracket relation

$$[\overline{\mathcal{X}}^m, \overline{\mathcal{X}}^k] = (k-m)\overline{\mathcal{X}}^{m+k-1}$$

for all $m, k \geq 0$ (cf. [33]). Here $\mathcal{X}^0$ is understood to be the string vector field $\mathcal{S}$. The relation between these vector fields and the Virasoro conjecture is explained in equation (4) and in more detail in [33].

Another sequence of vector fields is $T^n(\mathcal{X})$, $n \geq 0$, which satisfy the relation

$$[T^k(\mathcal{X}),\ T^m(\mathcal{X})] = (m-k)T^{m+k}(\mathcal{X})$$

for all $m, k \geq 0$. This relation follows from the following properties of the covariant differentiation

$$\nabla_{\mathcal{W}}\mathcal{X} = -Q(\mathcal{W}) + (b_1+1)\mathcal{W}, \quad \nabla_{\mathcal{V}}\, T^k(\mathcal{W}) = T^k(\nabla_{\mathcal{V}}\mathcal{W}) - T^{k-1}(V \circ W)$$

for all vector fields $\mathcal{V}$ and $\mathcal{W}$, as well as the following properties for operators T and Q

$$T^k Q T^m - T^m Q T^k = (m-k)T^{m+k}$$

for $m, k > 0$, and

$$(QT^k - T^k Q)(\mathcal{W}) = k\, T^k(\mathcal{W}) + T^{k-1}(\mathcal{X} \circ \mathcal{W})$$

for $k > 0$ and any vector field $\mathcal{W}$. The relation between vector fields $T^n(\mathcal{X})$ and the Virasoro conjecture is not clear at this moment.

4. Universal equations and spin curves

Universal equations also arise in the study of intersection theory on moduli spaces of spin curves. As mentioned above, the generating function for intersection numbers on the moduli spaces of stable curves is a τ-function of KdV hierarchy by Kontsevich–Witten theorem. In [46], Witten also proposed an algebraic geometric way to produce a τ-function for more general Gelfand–Dickey hierarchies by considering intersection numbers on the moduli spaces of spin curves. Assume that $r \geq 2$ is an integer and $\boldsymbol{m} = (m_1, \ldots, m_k)$ is a collection of integers with $0 \leq m_i \leq r-2$. When r is prime, an r-spin structure of type $\boldsymbol{m}$ over a smooth stable curve $(C; x_1, \ldots, x_k)$ is a line bundle $L \longrightarrow C$ together with an isomorphism $L^{\otimes r} \longrightarrow \omega\big(-\sum_{i=1}^k m_i x_i\big)$ where ω is the canonical line bundle over C. For degree reasons such a line bundle exists only if $\big(2g-2-\sum_{i=1}^k m_i\big)/r \in \mathbb{Z}$ where g is the genus of C. If r is not prime, then all d-th roots of $\omega\big(-\sum_{i=1}^k m_i x_i\big)$ should be considered for all d which divides r. If C is not smooth the definition of an r-spin structure is more involved (cf. [23] for details). A stable curve with an r-spin structure is called an r-spin curve. Let $\bar{\mathcal{M}}_{g,k}^{1/r}(\boldsymbol{m})$ be the moduli space of all genus-g r-spin curves of type $\boldsymbol{m}$ and with k-marked points. Let Ψ_i be the first Chern class of the line bundle over $\bar{\mathcal{M}}_{g,k}^{1/r}(\boldsymbol{m})$ whose geometric fiber over each r-spin curve is given by $T_{x_i}^* C$.

Let $e_0, \ldots, e_{r-2}$ be some abstract symbols. Define

$$\langle \tau_{n_1}(e_{m_1}) \ldots \tau_{n_k}(e_{m_k}) \rangle_{g,r} := r^{1-g} \int_{\overline{\mathcal{M}}_{g,k}^{1/r}(\boldsymbol{m})} c_{g,n}^{1/r}(\boldsymbol{m}) \cup \Psi_1^{n_1} \cup \cdots \cup \Psi_k^{n_k}$$

where $c_{g,n}^{1/r}(\boldsymbol{m})$ is a rational cohomology class on $\overline{\mathcal{M}}_{g,n}^{1/r}(\boldsymbol{m})$ of degree

$$\frac{2}{r}\Big\{(r-2)(g-1) + \sum_{i=1}^{n} m_i\Big\}.$$

The Poincaré dual of $c_{g,n}^{1/r}(\boldsymbol{m})$ corresponds to the virtual fundamental class in the Gromov–Witten theory. The genus-g generating function $\Phi_{g,r}$ for such numbers is then a function of parameters t_n^m with $0 \le m \le r-2$ and $n \in \mathbb{Z}_{\ge 0}$. Let $\Phi_r = \sum_{g=0}^{\infty} \Phi_{g,r}$. The *generalized Witten conjecture* predicts that $\exp(\Phi_r)$ is a τ-function of r-th KdV hierarchy (also called the Gelfand–Dickey hierarchy). More precisely, consider a differential operator

$$L = D^r - \sum_{i=0}^{r-2} u_i(x)\ D^i, \quad \text{where } D := \frac{\sqrt{-1}}{\sqrt{r}} \frac{\partial}{\partial x}.$$

An r-th root of L is a pseudo-differential operator of the form

$$L^{1/r} = D + \sum_{i>0} w_i(x) D^{-i}$$

whose r-th power is L and w_i are differential polynomials in $u_0, \ldots u_{r-2}$. Assume that L also depends on infinitely many parameters t_n^m with $0 \le m \le r-2$ and $n \ge 0$. The variable x is usually identified with t_0^0. We say that L satisfies the *r-th KdV hierarchy* if

$$\sqrt{-1}\ \frac{\partial L}{\partial t_n^m} = \frac{k_{n,m}}{\sqrt{r}} \big[\big(L^{n+\frac{m+1}{r}}\big)_+,\ L \big] \tag{5}$$

for all m and n, where

$$k_{n,m} = \frac{(-1)^n r^{n+1}}{(m+1)(r+m+1)\ldots(nr+m+1)}$$

and $\big(L^{n+\frac{m+1}{r}}\big)_+$ is a differential operator obtained by discarding all negatives powers of D in $L^{n+\frac{m+1}{r}}$. The *generalized Witten conjecture* says that there exists an operator L which satisfies the r-th KdV hierarchy such that

$$\frac{\partial^2 \Phi_r}{\partial t_0^0\, \partial t_n^m} = -k_{n,m} \operatorname{Res} \big(L^{n+\frac{m+1}{r}}\big) \tag{6}$$

for all m and n. Here "Res" means taking the coefficient of D^{-1}.

In some sense the generating functions $\Phi_{g,r}$ behaves like the generating function F_g for Gromov–Witten invariants of a fictitious manifold of rational dimension $2(r-2)/r$ and with the first Chern class equal to 0. Each abstract symbol e_m plays the role of a cohomology class of this fictitious manifold with rational degree $2m/r$. One can use the bilinear form defined by $(e_i,\ e_j) := \delta_{i+j,r-2}$ as a substitute for the Poincaré pairing. In this way, all structures of Gromov–Witten invariants mentioned in previous sections can make sense for intersection numbers on the moduli spaces of spin curves. In particular, $\Phi_{g,r}$ also satisfies the string equation. Together with the initial condition $\Phi(0) = 0$, the generalized Witten conjecture and the string equation completely determine the intersection numbers $\langle \tau_{n_1}(e_{m_1}) \dots \tau_{n_k}(e_{m_k}) \rangle_{g,r}$. It is also well known that the generalized Witten conjecture and the string equation are equivalent to Virasoro constraints formulated in a similar way (See for example [1]). There is also a canonical map $\overline{\mathcal{M}}^{1/r}_{g,k}(\boldsymbol{m}) \longrightarrow \overline{\mathcal{M}}_{g,k}$ which forgets spin structures on underlying stable curves. Therefore the functions $\Phi_{g,r}$ also satisfy universal equations for Gromov–Witten invariants which are obtained from tautological relations on $\overline{\mathcal{M}}_{g,k}$ (cf. [23]). In particular, a proof for the Virasoro conjecture only using universal equations also gives a proof to the generalized Witten conjecture.

On the other hand, if we assume that the generalized Witten's conjecture is true, we should also get a lot of information for universal equations for Gromov–Witten invariants. For this purpose it is desirable to write equations (5) and (6) in a form closer to universal equations. Analogous to Gromov–Witten theory, we define $\langle\!\langle\, \mathcal{W}_1 \ \dots \ \mathcal{W}_k \,\rangle\!\rangle_g$ and $\langle\!\langle\, \mathcal{W}_1 \ \dots \ \mathcal{W}_k \,\rangle\!\rangle$ as in equation (1) with F_g replaced by $\Phi_{g,r}$ and Φ_r respectively. We also define the grading operator G as a linear operator on the space of vector fields on the big phase space by

$$G(\tau_n(e_m)) := (n + b_m)\, \tau_n(e_m), \quad \text{with } b_m := \frac{m+1}{r}.$$

For any vector field $\mathcal{W}$ on the big phase space define

$$\tilde{T}(\mathcal{W}) := \tau_+(\mathcal{W}) - \sum_{m=0}^{r-2} \langle\!\langle\, \mathcal{W}\, e_m \,\rangle\!\rangle\, e_{r-2-m} \quad \text{and} \quad \tilde{R}(\mathcal{W}) := G\, \tilde{T}(\mathcal{W}).$$

The difference of operator $\tilde{T}$ and the operator T is that the coefficient of e_{r-2-m}, i.e. $\langle\!\langle\, \mathcal{W}\, e_m \,\rangle\!\rangle$, contains information for all genera, not only genus-0. Note that we think of $c_1 = 0$ for the theory of spin curves, so $\tilde{R}$ is a direct analogue of R with T replaced by $\tilde{T}$. In particular, we have

$$\tilde{R}(\tau_n(e_m)) = \left(n + 1 + \frac{m+1}{r}\right) \tau_{n+1}(e_m) - \sum_{j=0}^{r-2} \frac{r-1-j}{r} \langle\!\langle\, \tau_n(e_m)\, e_j \,\rangle\!\rangle\, e_{r-2-j}.$$

For any vector field $\mathcal{W}$ on the big phase space and $0 \le m \le r-2$, define

$$\begin{aligned}\Omega_m(\mathcal{W}) := {} & \langle\!\langle \tilde{R}(\mathcal{W})\, e_0\, e_m \rangle\!\rangle - \sum_{i=0}^{r-2} b_i^2 b_{r-2-i} \langle\!\langle \mathcal{W}\, e_0\, e_m\, e_i\, e_{r-2-i} \rangle\!\rangle \\ & - \sum_{i=0}^{r-2} \frac{b_{r-2-i}+b_m}{2}\ \langle\!\langle \mathcal{W}\, e_0\, e_i \rangle\!\rangle \langle\!\langle e_{r-2-i}\, e_m \rangle\!\rangle \\ & - \sum_{i=0}^{r-2} \frac{b_{r-2-i}+b_0}{2}\ \langle\!\langle \mathcal{W}\, e_m\, e_i \rangle\!\rangle \langle\!\langle e_{r-2-i}\, e_0 \rangle\!\rangle . \end{aligned} \tag{7}$$

For each r, the generalized Witten conjecture can be reformulated in the form

$$\Omega_m(\mathcal{W}) = \cdots \tag{8}$$

for all vector field $\mathcal{W}$, where the right-hand side of this equation is an expression each term of which involves only $\mathcal{W}$ and at least 6 primary vector fields. For example, when $r = 2$, Witten's formulation of his KdV conjecture in [44] is equivalent to

$$\Omega_0(\mathcal{W}) = 0 \tag{9}$$

for all vector field $\mathcal{W}$. Shadrin [43] shows that the generalized Witten conjecture for $r = 3$ also implies a formula which is equivalent to equation (9). In [38], it is proved that the generalized Witten conjecture for $r = 3$ also implies that

$$\Omega_1(\mathcal{W}) = \frac{1}{108}\left\{ -\langle\!\langle \mathcal{W}\, e_0^2 \rangle\!\rangle \langle\!\langle e_0^4 \rangle\!\rangle - \langle\!\langle \mathcal{W}\, e_0^3 \rangle\!\rangle \langle\!\langle e_0^3 \rangle\!\rangle + \langle\!\langle \mathcal{W}\, e_0^4 \rangle\!\rangle \langle\!\langle e_0^2 \rangle\!\rangle \right\}$$

for all vector field $\mathcal{W}$. In fact, this equation and equation (9) together are equivalent to the generalized Witten conjecture for $r = 3$. In this formula and the formula below e_i^k does not mean the quantum power of e_i. It simply means e_i repeating k times.

It seems quite hard to give a general formula for the right-hand side of equation (8) for all r, even for the special case $m = 0$. Equation (9) only holds for $r = 2$ and $r = 3$. It does not hold when $r > 3$. For example, when $r = 4$, we have

$$\Omega_0(\mathcal{W}) = \frac{1}{192}\left\{ \langle\!\langle \mathcal{W}\, e_0^2 \rangle\!\rangle \langle\!\langle e_0^4 \rangle\!\rangle + \langle\!\langle \mathcal{W}\, e_0^3 \rangle\!\rangle \langle\!\langle e_0^3 \rangle\!\rangle - \langle\!\langle \mathcal{W}\, e_0^4 \rangle\!\rangle \langle\!\langle e_0^2 \rangle\!\rangle \right\}$$

as a consequence of the generalized Witten conjecture (cf. [38]). However, from all examples computed in [38], we expect that at the origin of the big phase space $\Omega_m(\mathcal{W}) = 0$ for all $r \ge 2$ and all vector fields $\mathcal{W}$ on the big phase space. Combining with the string equation, the equation $\Omega_0(\mathcal{W}) = 0$ at the origin gives the following recursion formula for intersection numbers:

$$\left(n+1+\frac{m+2}{r}\right)\langle \tau_n(e_m) \rangle = \sum_{j=0}^{r-2} \frac{(j+1)^2(r-1-j)}{r^3} \langle \tau_{n-1}(e_m)\, e_j\, e_{r-2-j} \rangle$$

for all r, m and n. This formula has been checked for $r \leq 7$ in [38] using the generalized Witten conjecture. In particular, we used this formula computed $\langle \tau_n(e_m) \rangle_{3,r}$. The results match with the computations using the genus-3 topological recursion relation in [25].

Starting from any vector field $\mathcal{W}$ on the big phase space, the above equations obtained from the generalized Witten conjecture are recursion relations involving $\tilde{R}(\mathcal{W})$. One can also write them as recursion relations involving $R(\mathcal{W})$ in more complicated forms. In comparison, universal equations obtained from tautological relations are recursion relations involving $T(\mathcal{W})$. In this sense the recursion relations from the generalized Witten conjecture are closer to the Virasoro constraints rather than universal equations. It would be interesting to find out relations between these two types of recursion relations. In [44], Witten showed how to check the compatibility of the equation (9) for $r = 2$ and the topological recursion relations of genus-0 and genus-1. He employed the constitutive relations which were obtained in [9] using genus-0 and genus-1 topological recursion relations. So far there is no analogue for the constitutive relations when the genus is bigger than one. This makes it harder to check the compatibility even for the case when $r = 2$ if the genus is bigger than 1. We believe that it is very important to understand the relations between the generalized Witten conjecture (or τ-functions of Gelfand–Dickey hierarchies) and universal equations of Gromov–Witten invariants. Such relations should be crucial in understanding the structures of the complicated systems of universal equations.

References

[1] Adler, M., and Van Moerbeke, P., A matrix integral solution to two-dimensional W_p Gravity. *Comm. Math. Phys.* **147** (1992), 25–56.

[2] Arcara, D., and Lee, Y.-P., Tautological equation in $\overline{\mathcal{M}}_{3,1}$ via invariance conjectures. math.AG/0503184.

[3] Behrend, K., and Fantechi, B., The intrinsic normal cone. *Invent. Math.* **128** (1997), 45–88.

[4] Belorousski, P., and Pandharipande, R., A descendent relation in genus 2. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **29** (2000), 171–191.

[5] Bertram, A., Ciocan-Fontanine, I., and Kim, B., Two Proofs of a conjecture of Hori and Vafa. *Duke Math. J.* **126** (1) (2005), 101–136.

[6] Cox, D., and Katz, S., *Mirror symmetry and algebraic geometry*. Math. Surveys Monogr. 68, Amer. Mathem. Soc., Providence, RI, 1999.

[7] Deligne, P., and Mumford, D., The irreducibility of the space of curves of given genus. *Inst. Hautes Études Sci. Publ. Math.* **36** (1969), 75–109.

[8] Deligne, P., and Rapoport, M. D., Les schémas de modules de courbes elliptiques. In *Modular functions of one variable, II* (Antwerp 1972), Lecture Notes in Math. 349, Springer-Verlag, Berlin 1973, 143–316.

[9] Dijkgraaf, R., and Witten, E., Mean field theory, topological field theory, and multimatrix models. *Nucl. Phys. B* **342** (1990), 486–522.

[10] Dubrovin, B., Geometry of 2D topological field theories. In *Integrable Systems and Quantum Groups*, Lectures Notes in Math. 1620, Springer-Verlag, Berlin 1996, 120–348.

[11] Dubrovin, B., and Zhang, Y., Bihamiltonian hierarchies in 2D topological field theory at one-loop approximation. *Comm. Math. Phys.* **198** (2) (1998), 311–361.

[12] Dubrovin, B., and Zhang, Y., Frobenius manifolds and Virasoro constraints. *Selecta Math.* (N.S.) **5** (1999), 423–466.

[13] Dubrovin, B., and Zhang, Y., Normal forms of hierarchies of integrable PDEs, Frobenius manifolds and Gromov-Witten invariants. math.DG/0108160.

[14] Eguchi, T., Hori, K., and Xiong, C., Quantum Cohomology and Virasoro Algebra. *Phys. Lett. B* **402** (1997), 71–80.

[15] Faber, C., and Pandharipande, R., Relative maps and tautological classes. *J. Eur. Math. Soc.* **7** (1) (2005), 13–49.

[16] Getzler, E., Intersection theory on $\bar{M}_{1,4}$ and elliptic Gromov-Witten Invariants. *J. Amer. Math. Soc.* **10** (1997), 973–998.

[17] Getzler, E., Topological recursion relations in genus 2. In *Integrable systems and algebraic geometry* (Kobe/Kyoto, 1997), World Sci. Publishing, River Edge, NJ, 1998, 73–106.

[18] Getzler, E., The Virasoro conjecture for Gromov-Witten invariants. In *Algebraic Geometry: Hirzebruch 70* (Warsaw, 1998), Contemp. Math. 241, Amer. Math. Soc., Providence, RI, 1999, 147–176.

[19] Givental, A., Semisimple Frobenius structures at higher genus. *Internat. Math. Res. Notices* **23** (2001), 1265–1286.

[20] Givental, A., Gromov-Witten invariants and quantization of quadratic hamiltonians. *Moscow Math. J.* **1** (4) (2001), 551–568.

[21] Givental, A., Symplectic geometry of Frobenius structures. In *Frobenius manifolds*, Aspects Math. E36, Vieweg, Wiesbaden 2004, 91–112.

[22] Ionel, E., Topological recursive relations in $H^{2g}(\mathcal{M}_{g,n})$. *Invent. Math.* **148** (3) (2002), 627–658.

[23] Jarvis, T., Kimura, T., and Vaintrob, A., Moduli Spaces of Higher Spin Curves and Integrable Hierarchies. *Compositio Math.* **126** (2) (2001), 175–212.

[24] Joe, D., and Kim., B., Equivariant mirrors and the Virasoro conjecture for flag manifolds. *Internat. Math. Res. Notices* **15** (2003), 859–882.

[25] Kimura, T., and Liu, X., A genus-3 topological recursion relation. *Comm. Math. Phys.* **262** (3) (2006), 645–661.

[26] Kontsevich, M., Intersection theory on the moduli space of curves and the matrix airy function. *Comm. Math. Phys.* **147** (1992), 1–23 .

[27] Kontsevich, M., and Manin, Y., Gromov-Witten classes, quantum cohomology, and enumerative geometry. *Comm. Math. Phys.* **164** (1994), 525–562.

[28] Kontsevich, M., and Manin, Y., Relations between the correlators of the topological sigma-model coupled to gravity. *Comm. Math. Phys.* **196** (1998), 385–398.

[29] Lee, Y.-P., Witten's conjecture, Virasoro conjecture, and invariance of tautological equations. math.AG/0311100.

[30] Li, J., and Tian, G., Virtual moduli cycles and Gromov-Witten invariants of general symplectic manifolds. In *Topics in symplectic* 4*-manifolds* (Irvine, CA, 1996), First Int. Press Lect. Ser., I, Internat. Press, Cambridge, MA, 1998, 47–83.

[31] Li, J., and Tian, G., Virtual moduli cycles and Gromov-Witten invariants of algebraic varieties. *J. Amer. Math. Soc.* **11** (1998), 119–174.

[32] Liu, X., Elliptic Gromov-Witten invariants and Virasoro conjecture. *Comm. Math. Phys.* **216** (2001), 705–728.

[33] Liu, X., Quantum product on the big phase space and Virasoro conjecture. *Adv. Math.* **169** (2002), 313–375.

[34] Liu, X., Quantum product, topological recursion relations, and the Virasoro conjecture. To appear in Proceedings of Mathematical Society of Japan - 9th International Research Institute on *Integrable Systems in Differential Geometry*, Tokyo, Japan.

[35] Liu, X., Relations among universal equations for Gromov-Witten invariants. In *Frobenius manifolds*, Aspects Math. E36, Vieweg, Wiesbaden 2004, 169–180.

[36] Liu, X., Idempotents on the big phase space. To appear in the *Proceedings of the conference on Gromov-Witten Theory of Spin Curves and Orbifolds* (San Francisco 2003); math.DG/0310409.

[37] Liu, X., Genus-2 Gromov-Witten invariants for manifolds with semisimple quantum cohomology. *Amer. J. Math.*, to appear; math.DG/0310410.

[38] Liu, X. Recursion formulas for intersection numbers on the moduli spaces of spin curves. In preparation.

[39] Liu, X., and Tian, G., Virasoro constraints for quantum cohomology. *J. Differential Geom.* **50** (1998), 537–591.

[40] Mumford, D., Towards an enumerative geometry of the moduli space of curves. In *Arithmetic and geometry*, Vol. II, Progr. Math. 36, Birkhäuser, Boston 1983, 271–328.

[41] Okounkov, A., and Pandharipande, R., Virasoro constraints for target curves. math.AG/0308097.

[42] Ruan, Y., and Tian, G., A mathematical theory of quantum cohomology. *J. Differential Geom.* 42 (1995), 259–367.

[43] Shadrin, S., Geometry of meromorphic functions and intersections on moduli spaces of curves. *Internat. Math. Res. Notices* **2003** (38) (2003), 2051–2094,

[44] Witten, E., Two dimensional gravity and intersection theory on Moduli space. In *Surveys in differential geometry* 1, Supplement to *J. Differential Geom.* (1991), 243–310.

[45] Witten, E., On the Kontsevich model and other models of two dimensional gravity. In *Proceedings of the XXth international conference on differential geometric methods in theoretical physics* (New York, 1991), World Sci. Publishing, River Edge, NJ, 1992, 176–216.

[46] Witten, E., Algebraic geometry associated with matrix models of two dimensional gravity. *Topological methods in modern mathematics* (Stony Brook, NY, 1991), Publish or Perish, Houston, TX, 1993, 235–269.

Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556, U.S.A.
E-mail: xliu3@nd.edu

Extremal metrics and stabilities on polarized manifolds

Toshiki Mabuchi*

Abstract. The Hitchin–Kobayashi correspondence for vector bundles, established by Donaldson, Kobayashi, Lübke, Uhlenbeck and Yau, states that an indecomposable holomorphic vector bundle over a compact Kähler manifold is stable in the sense of Takemoto–Mumford if and only if the vector bundle admits a Hermitian-Einstein metric. Its manifold analogue known as Yau's conjecture, which originated from Calabi's conjecture, asks whether "stability" and "existence of extremal metrics" for polarized manifolds are equivalent. In this note the recent progress of this subject, by Donaldson, Tian and our group, together with its relationship to algebraic geometry will be discussed.

Mathematics Subject Classification (2000). Primary 32Q15; Secondary 53C21.

Keywords. Asymptotic Bergman kernel, balanced metrics, Calabi's conjecture, Catlin–Lu–Tian–Zelditch's theorem, Chow–Mumford stable, extremal Kähler metrics, Futaki's character, GIT (geometric invariant theory), geometry of Kähler potentials, Gieseker's stability for surfaces of general type, Hilbert–Mumford stable, Hitchin–Kobayashi correspondence, Kähler–Einstein metrics, K-energy, K-stable, Matsushima's obstruction, stability theorems by Donaldson and Tian, Takemoto–Mumford stable, Yau's conjecture.

1. Introduction

Let M be a compact complex connected manifold. As an introduction to our subject we recall the following well-known conjecture of Calabi [5]:

Conjecture. (i) If $c_1(M)_{\mathbb{R}} < 0$, then M admits a unique Kähler–Einstein metric ω such that $\mathrm{Ric}(\omega) = -\omega$.

(ii) If $c_1(M)_{\mathbb{R}} = 0$, then each Kähler class on M admits a unique Kähler–Einstein metric ω such that $\mathrm{Ric}(\omega) = 0$.

(iii) For $c_1(M)_{\mathbb{R}} > 0$, find a suitable condition for M to admit a Kähler–Einstein metric ω such that $\mathrm{Ric}(\omega) = \omega$.

Based on the pioneering works of Calabi [6], [7] and Aubin [1], a complete affirmative answer to (i) and (ii) was given by Yau [49] by solving systematically certain complex Monge–Ampère equations. However for (iii), sufficient conditions are known only partially by Siu [41], Nadel [37], Tian [42], [44], Tian and Yau [46], Wang and Zhu [52], where some necessary conditions were formulated as obstructions by

*Special thanks are due to Professors K. Hirachi and G. Komatsu for valuable suggestions.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Futaki [18] and Matsushima [34] (see also Lichnerowicz [23]). It is (iii) that mainly motivates our studies of stabilities and extremal metrics, while an analysis of destabilizing phenomena caused by the non-existence of Kähler–Einstein metrics allows us to obtain very nice by-products, such as Nadel's vanishing theorem [37] via the use of multiplier ideal sheaves.

2. Stability for manifolds in algebraic geometry

In Mumford's GIT [36], moduli spaces of algebraic varieties are constructed via the theory of invariants, where varieties are described by numerical data modulo actions of reductive algebraic groups. Then, roughly speaking, stable points are those away from very bad points in moduli spaces. For a precise definition consider a representation of a reductive algebraic group G on a complex vector space W. Let $w \in W$. We denote by G_w the isotropy subgroup of G at w.

Definition 2.1. A point $w \in W$ is said to be *stable* (resp. *properly stable*) if the orbit $G \cdot w$ is closed in W (resp. $G \cdot w$ is closed in W with $|G_w| < \infty$).

For moduli spaces of polarized varieties the Chow–Mumford stability and the Hilbert–Mumford stability are known. In what follows, by a polarized manifold (M, L), we mean a very ample holomorphic line bundle L over a nonsingular projective algebraic variety M defined over $\mathbb{C}$, where the arguments in this section have nothing to do with the nonsingularity of M. Now for a polarized manifold (M, L) put $n := \dim M$ and let m be a positive integer. Then associated to the complete linear system $|L^m|$ we have the Kodaira embedding

$$\iota_m : M \hookrightarrow \mathbb{P}^*(V_m),$$

where $\mathbb{P}^*(V_m)$ denotes the set of all hyperplanes in $V_m := H^0(M, \mathcal{O}(L^m))$ through the origin. Let d_m be the degree of $\iota_m(M)$ in the projective space $\mathbb{P}^*(V_m)$. Put $G_m := \mathrm{SL}_{\mathbb{C}}(V_m)$ and $W_m := \{S^{d_m}(V_m)\}^{\otimes n+1}$, where $S^{d_m}(V_m)$ is the d_m-th symmetric tensor product of the space V_m. Take an element $M_m \neq 0$ in W_m^* such that the associated element $[M_m]$ in $\mathbb{P}^*(W_m)$ is the Chow point of the irreducible reduced algebraic cycle $\iota_m(M)$ on $\mathbb{P}^*(V_m)$. For the natural action of G_m on W_m^* we now apply Definition 2.1 to $G = G_m$ and $W = W_m$:

Definition 2.2. (a) (M, L^m) is said to be *Chow–Mumford stable* (resp. *Chow–Mumford properly stable*) if M_m in W_m^* is stable (resp. properly stable).

(b) (M, L) is said to be *asymptotically Chow–Mumford stable* (resp. *asymptotically Chow–Mumford properly stable*) if, for $m \gg 1$, (M, L^m) is Chow–Mumford stable (resp. Chow–Mumford properly stable).

Let m and k be positive integers. Then the kernel $I_{m,k}$ of the natural homomorphism of $S^k(V_m)$ to $V_{mk} := H^0(M, \mathcal{O}(L^{mk}))$ is the homogeneous ideal of degree k

defining M in $\mathbb{P}^*(V_m)$. Put $N_{mk} := \dim V_{mk}$ and $\gamma_{m,k} := \dim I_{m,k}$. Then $\wedge^{\gamma_{m,k}} I_{m,k}$ is a complex line in the vector space $W_{m,k} := \wedge^{\gamma_{m,k}}(S^k(V_m))$. Take an element $w_{m,k} \neq 0$ in $\wedge^{\gamma_{m,k}} I_{m,k}$. For the natural action of $G_m := \mathrm{SL}_{\mathbb{C}}(V_m)$ on $W_{m,k}$ we apply Definition 2.1 to $G = G_m$ and $W = W_{m,k}$:

Definition 2.3. (a) (M, L^m) is said to be *Hilbert–Mumford stable* (resp. *Hilbert–Mumford properly stable*) if $w_{\ell,k} \in W_{\ell,k}$ is stable (resp. properly stable).

(b) (M, L) is said to be *asymptotically Hilbert–Mumford stable* (resp. *asymptotically Hilbert–Mumford properly stable*) if, for all $m \gg 1$, (M, L^m) is Hilbert–Mumford stable (resp. Hilbert–Mumford properly stable).

A result of Fogarty [17] shows that if (M, L^m) is Chow–Mumford stable, then (M, L^m) is also Hilbert–Mumford stable. Though the converse has been unknown, the relationship between these two stabilities is now becoming clear (cf. [33]).

Stability for manifolds is an important subject in moduli theories of algebraic geometry. Recall, for instance, the following famous result of Mumford [35]:

Fact 1. *If L is an ample line bundle of degree $d \geq 2g + 1$ over a compact Riemann surface C of genus $g \geq 1$, then (C, L) is Chow–Mumford properly stable.*

For the pluri-canonical bundles $K_M^{\otimes m}$ on $M, m \gg 1$, using the asymptotic Hilbert–Mumford stability, Gieseker [20] generalized this result to the case where M is a surface of general type. For higher dimensions a stability result by Viehweg [48] is known in the case where the canonical bundle K_M is semipositive. However, for both the results of Gieseker and of Viehweg the proof of stability is fairly complicated, while the underlying manifold (or orbifold) admits a Kähler–Einstein metric.

3. The Hitchin–Kobayashi correspondence and its manifold analogue

For a holomorphic vector bundle E over an n-dimensional compact Kähler manifold (M, ω) we say that E is *Takemoto–Mumford stable* if

$$\frac{\int_M c_1(\mathcal{S})\omega^{n-1}}{\mathrm{rk}(\mathcal{S})} < \frac{\int_M c_1(E)\omega^{n-1}}{\mathrm{rk}(E)}$$

for every coherent subsheaf $\mathcal{S}$ of $\mathcal{O}(E)$ satisfying $0 < \mathrm{rk}(\mathcal{S}) < \mathrm{rk}(E)$. Recall the following Hitchin–Kobayashi correspondence for vector bundles:

Fact 2. *An indecomposable holomorphic vector bundle E over M is Takemoto–Mumford stable if and only if E admits a Hermitian-Einstein metric.*

This fact was established in 1980s by Donaldson [13], Kobayashi [22], Lübke [25], Uhlenbeck and Yau [47]. As a manifold analogue of this conjecture we can naturally ask whether the following conjecture (known as Yau's conjecture) is true:

Conjecture. The polarization class of (M, L) admits a Kähler metric of constant scalar curvature (or more generally an extremal Kähler metric) if and only if (M, L) is asymptotically stable in a certain sense of GIT.

For the "only if" part of this conjecture, the first breakthrough was made by Tian [45]. By introducing the concept of K-stability, he gave an answer to the "only if" part for Kähler–Einstein manifolds, and showed that some Fano manifolds without nontrivial holomorphic vector fields admit no Kähler–Einstein metrics. A remarkable progress was made by Donaldson [14] who showed the Chow–Mumford stability for a polarized Kähler manifold (M, ω) of constant scalar curvature essentially when the connected linear algebraic part H of the group $\mathrm{Aut}(M)$ of holomorphic automorphisms of M is semisimple. In the present paper we shall show how Donaldson's work is generalized to extremal Kähler cases without any assumption on H (see also [31], [32]). The relationship between this generalization and a recent result by Chen and Tian [12] will be treated elsewhere.

4. The asymptotic Bergman kernel

For a polarized manifold (M, L) take a Hermitian metric h for L such that $\omega := c_1(L; h)$ is a Kähler form. Define a Hermitian pairing on $V_m := H^0(M, \mathcal{O}(L^m))$ by

$$\langle \sigma_1, \sigma_2 \rangle_{L^2} := \int_M (\sigma_1, \sigma_2)_h \, \omega^n, \quad \sigma_1, \sigma_2 \in V_m,$$

where $(\ ,\ )_h$ denotes the pointwise Hermitian pairing by h^m for sections for L^m. For an orthonormal basis $\{\sigma_1, \sigma_2, \ldots, \sigma_{N_m}\}$ of V_m we put

$$B_{m,\omega} := \frac{n!}{m^n}(|\sigma_1|_h^2 + |\sigma_2|_h^2 + \cdots + |\sigma_{N_m}|_h^2), \tag{1}$$

where $|\sigma|_h^2 := (\sigma, \sigma)_h$ for $\sigma \in V_m$. This $B_{m,\omega}$ is called the m-th *Bergman kernel* for (M, ω) (cf. Tian [43], Zelditch [50], Catlin [4]), where we consider the asymptotic behavior of $B_{m,\omega}$ as $m \to \infty$. Note that $B_{m,\omega}$ depends only on (m, ω) and is independent of the choice of both h and the orthonormal basis for V_m. Next, for

$$D := \{\ell \in L^* ;\ |\ell|_h < 1\}$$

the boundary $X := \partial D = \{\ell \in L^*; |\ell|_h = 1\}$ over M is an S^1-bundle. Let $\mathrm{pr}\colon X \to M$ be the natural projection. We now consider the Szegö kernel

$$S_\omega := S_\omega(x, y)$$

for the projection of $L^2(X)$ onto the Hardy space $L^2(X) \cap \Gamma(D, \mathcal{O})$ of boundary values of holomorphic functions on D. Then for each positive integer m the corresponding

m-th Bergman kernel $B_{m,\omega}$ for the Kähler manifold (M, ω) is characterized as the Fourier coefficient

$$\mathrm{pr}^* B_{m,\omega} := \frac{n!}{m^n} \int_{S^1} e^{-im\theta} S_\omega(e^{i\theta}x, x)\, d\theta.$$

Now the Bergman kernel is defined not only for positive integers m but also for complex numbers ξ as follows. To see the situation, we first consider the case where M is a single point. Then in place of $S_\omega(e^{i\theta}x, x)$ consider a smooth function $S = S(\theta)$ on $S^1 := \mathbb{R}/2\pi\mathbb{Z}$ for simplicity. The associated Fourier coefficient B_m is

$$B_m = \int_{S^1} e^{-im\theta} S(\theta)\, d\theta$$

for each integer $m \neq 0$. Then for open intervals $I_1 := (-3\pi/4, 3\pi/4)$ and $I_2 = (\pi/4, 7\pi/4)$ in $\mathbb{R}$ we choose the open cover

$$S^1 = U_1 \cup U_2$$

where $U_1 := I_1 \bmod 2\pi$ and $U_2 := I_2 \bmod 2\pi$. By choosing a partition of unity subordinate to this open cover we write

$$\rho_1(\theta) + \rho_2(\theta) = 1, \quad \theta \in S^1,$$

where $\rho_\alpha \in C^\infty(S^1)_\mathbb{R}$, $\alpha = 1, 2$, are functions ≥ 0 satisfying $\mathrm{Supp}(\rho_\alpha) \subset U_\alpha$. For the coordinate $\tilde{\theta}$ for $\mathbb{R}$, writing $\tilde{\theta} \bmod 2\pi$ as θ, define $\tilde{\rho}_\alpha \in C^\infty(\mathbb{R})_\mathbb{R}$, $\alpha = 1, 2$, by

$$\tilde{\rho}_\alpha(\tilde{\theta}) = \begin{cases} \rho_\alpha(\theta), & \tilde{\theta} \in I_\alpha \\ 0, & \tilde{\theta} \notin I_\alpha. \end{cases}$$

Then the Fourier transform $\mathcal{F}(S) = \mathcal{F}(S)(\xi)$ of S is an entire function in $\xi \in \mathbb{C}$ defined as the integral

$$\mathcal{F}(S)(\xi) = \int_\mathbb{R} e^{-i\xi\theta}\{\rho_1(\tilde{\theta}) + \rho_2(\tilde{\theta})\}S(\tilde{\theta})\, d\tilde{\theta}, \quad \xi \in \mathbb{C},$$

satisfying $\mathcal{F}(S)(m) = B_m$ for all integers m. Though $\mathcal{F}(S)$ may depend on the choice of the partition of unity, its restriction to $\mathbb{Z}$ is unique. This situation is easily understood for instance by the fact the functions $\mathcal{F}(S)(\xi)$ and $\mathcal{F}(S)(\xi) + \sin(\pi\xi)$ in ξ coincide on $\mathbb{Z}$.

Now the above process of generalization from the Fourier series to the Fourier transform is valid also for the case where the base space M is nontrivial. Actually, we define an entire function $B_{\xi,\omega}$ in $\xi \in \mathbb{C}$ by

$$\mathrm{pr}^* B_{\xi,\omega} := \frac{n!}{\xi^n} \int_\mathbb{R} e^{-i\xi\tilde{\theta}}\{\rho_1(\tilde{\theta}) + \rho_2(\tilde{\theta})\}S_\omega(e^{i\tilde{\theta}}x, x)\, d\tilde{\theta}.$$

By setting $q := \xi^{-1}$ we study the asymptotic behavior of $B_{\xi,\omega}$ as $q \to 0$ along the positive real line $\{q > 0\}$. Let σ_ω denote the scalar curvature of the Kähler manifold (M, ω). Then as in discrete cases by Tian [43], Zelditch [50], Catlin [4], the asymptotic expansion of $B_{\xi,\omega}$ in q yields

$$B_{\xi,\omega} = 1 + a_1(\omega)q + a_2(\omega)q^2 + \cdots, \quad 0 \le q \ll 1, \tag{2}$$

where $a_1(\omega) = \sigma_\omega/2$ by a result of Lu [24]. For more details of the expansion in discrete cases see also Hirachi [21].

5. Balanced metrics

Choose a Hermitian metric h for L such that $\omega := c_1(L; h)$ is a Kähler form. Then ω is called an m-th *balanced metric* (cf. [51], [26]) for (M, L) if $B_{m,\omega}$ is a constant function $(= C_m)$ on M. First put $q := 1/m$. By integrating (1) and (2) on the Kähler manifold (M, ω) we see that C_m is written as

$$C_m := \frac{n!}{m^n c_1(L)^n[M]} N_m = 1 + \frac{n}{2}\frac{c_1(M)c_1(L)^{n-1}[M]}{c_1(L)^n[M]} q + O(q^2), \quad 0 \le q \ll 1,$$

where the left-hand side is the Hilbert polynomial $P(m)$ for (M, L) divided by $m^n c_1(L)^n[M]/n!$. Hence it is easy to define C_ξ by setting

$$C_\xi = \frac{n!P(\xi)}{\xi^n c_1(L)^n[M]}, \quad \xi \in \mathbb{C}^*. \tag{3}$$

Now put $q := 1/\xi$. Define the modified Bergman kernel $\beta_{q,\omega}$ by

$$\beta_{q,\omega} := 2\xi\left(1 + \frac{2q}{3}\Delta_\omega\right)(B_{\xi,\omega} - C_\xi) = \sigma_\omega - \bar{\sigma}_\omega + O(q), \tag{4}$$

where the average $\bar{\sigma}_\omega$ of the scalar curvature σ_ω is $nc_1(M)c_1(L)^{n-1}[M]/c_1(L)^n[M]$ independent of the choice of ω in $c_1(L)_{\mathbb{R}}$. Then, for $\xi = m$, ω is an m-th balanced metric for (M, L) if and only if $\beta_{q,\omega}$ vanishes everywhere on M. Recall the following result by Zhang [51] (cf. [26]; see also [30]):

Fact 3. *(M, L^m) is Chow–Mumford stable if and only if (M, L) admits an m-th balanced metric.*

Consider the maximal connected linear algebraic subgroup H of $\mathrm{Aut}(M)$, so that the identity component of $\mathrm{Aut}(M)/H$ is an Abelian variety. Let us now choose an algebraic torus $T \cong (\mathbb{C}^*)^r$ in the connected component $Z^{\mathbb{C}}$ of the center of the reductive part $R(H)$ for H in the Chevalley decomposition

$$H = R(H) \ltimes H_u,$$

where H_u is the unipotent radical of H. Replacing L by its suitable positive multiple, we may assume that the H-action on M is lifted to a bundle action on L covering the H-action on M. For each character $\chi \in \mathrm{Hom}(T, \mathbb{C}^*)$ we set

$$W\chi := \{\sigma \in V_m\,;\ \sigma \cdot g = \chi(g)\sigma \text{ for all } g \in T\},$$

where $V_m \times T \ni (\sigma, g) \mapsto \sigma \cdot g$ is the right T-action on $V_m = H^0(M, \mathcal{O}(L^m))$ induced by the left H-action on L. Now we have characters $\chi_k \in \mathrm{Hom}(T, \mathbb{C}^*)$, $k = 1, 2, \ldots, r_m$, such that the vector space V_m is expressible as a direct sum

$$V_m = \bigoplus_{k=1}^{r_m} W\chi_k.$$

For the maximal compact torus T_c in T we may assume that both h and ω are T_c-invariant. Put $\mathbb{J}_m := \{1, 2, \ldots, N_m\}$, where $N_m := \dim V_m$. Choose an orthonormal basis $\{\sigma_1, \sigma_2, \ldots, \sigma_{N_m}\}$ for V_m such that all σ_j, $j \in \mathbb{J}_m$, belong to the union $\bigcup_{k=1}^{r_m} W\chi_k$. Hence there exists a map $\kappa : \mathbb{J}_m \to \{1, 2, \ldots, r_m\}$ satisfying

$$\sigma_j \in W\chi_{\kappa(j)}, \quad j \in \mathbb{J}_m.$$

Put $\mathfrak{t}_{\mathbb{R}} := i\mathfrak{t}_c$ for the Lie algebra $\mathfrak{t}_c$ of T_c where $i := \sqrt{-1}$. For each $\mathcal{Y} \in \mathfrak{t}_{\mathbb{R}}$, by setting $g := \exp(\mathcal{Y}/2)$, we put $h_g := h \cdot g$ for the natural T-action on the space of Hermitian metrics on L. Define the m-th weighted Bergman kernel $B_{m,\omega,\mathcal{Y}}$, twisted by $\mathcal{Y}$, for (M, ω) by setting

$$B_{m,\omega,\mathcal{Y}} := \frac{n!}{m^n} \sum_{j=1}^{N_m} |\sigma_j|^2_{h_g} = g^* \left\{ \frac{n!}{m^n} \sum_{j=1}^{N_m} \frac{|\sigma_j|^2_h}{|\chi_{\kappa(j)}(g)|^2} \right\}.$$

Then ω is called an *m-th T-balanced metric* on (M, L) if $B_{m,\omega,\mathcal{Y}}$ is a constant function $(= C_{m,\mathcal{Y}})$ on M for some $\mathcal{Y} \in \mathfrak{t}$. Consider the natural action of the group

$$G_m := \bigoplus_{k=1}^{r_m} \mathrm{SL}_{\mathbb{C}}(W\chi_k)$$

acting on $V_m = \bigoplus_{k=1}^{r_m} W\chi_k$ diagonally (factor by factor). We say that (M, L^m) is *Chow–Mumford T-stable* if the orbit $G_m \cdot M_m$ is closed in W_m^*. Note that (cf. [30])

Fact 4. *If M admits an m-th T-balanced metric on (M, L), then (M, L^m) is Chow–Mumford T-stable.*

We shall now extend $\{B_{m,\omega,\mathcal{Y}}\,;\ m = 1, 2, \ldots\}$ to $\{B_{\xi,\omega,\mathcal{Y}}\,;\ \xi \in \mathbb{C}\}$ in such a way that $B_{m,\omega,\mathcal{Y}}$ coincides with $B_{\xi,\omega,\mathcal{Y}|\xi=m}$ for all positive integers m. By the definition of $B_{m,\omega,\mathcal{Y}}$ (see also [30], p. 578) the equality $B_{m,\omega,\mathcal{Y}} = B_{m,\omega}(h_g/h)^m$ always holds. Hence we put

$$B_{\xi,\omega,\mathcal{Y}} := B_{\xi,\omega} \cdot (h_g/h)^{\xi} = B_{\xi,\omega} \exp\{\xi \log(h_g/h)\}, \quad \xi \in \mathbb{C}. \tag{5}$$

Once $B_{\xi,\omega,\mathcal{Y}}$ is defined, we can also define $C_{\xi,\mathcal{Y}}$, $\xi \in \mathbb{C}$ in such a way that $C_{m,\mathcal{Y}}$ coincides with $C_{\xi,\mathcal{Y}_{|\xi=m}}$. Actually, we put

$$C_{\xi,\mathcal{Y}} := \int_M \frac{B_{\xi,\omega,\mathcal{Y}}}{c_1(L)^n[M]}\, g^*\omega^n. \tag{6}$$

6. A simple heuristic proof of Donaldson's theorem

In this section we shall show that a heuristic application of the implicit function theorem simplifies the proof of Donaldson's theorem [14]. Fix a Kähler metric ω_0 in $c_1(L)_{\mathbb{R}}$ of constant scalar curvature. Assume that the group H in the previous section is trivial. For each Kähler metric ω in $c_1(L)_{\mathbb{R}}$ we can associate a unique real-valued smooth function φ on M such that

$$\omega = \omega_0 + \sqrt{-1}\partial\bar{\partial}\varphi$$

with normalization condition $\int_M \varphi\, \omega_0^n = 0$. For an arbitrary nonnegative integer k and a real number α satisfying $0 < \alpha < 1$, we more generally consider the case where $\varphi \in C^{k+4,\alpha}(M)_{\mathbb{R}}$, so that ω is a $C^{k+2,\alpha}$ Kähler metric on M. The Fréchet derivative $D_\omega\sigma_\omega$ at $\omega = \omega_0$ of the scalar curvature function $\omega \mapsto \sigma_\omega$ is given by

$$\left\{(D_\omega\sigma_\omega)(\sqrt{-1}\partial\bar{\partial}\varphi)\right\}_{|\omega=\omega_0} = \lim_{\varepsilon\to 0} \frac{\sigma_{\omega_0+\sqrt{-1}\varepsilon\partial\bar{\partial}\varphi} - \sigma_{\omega_0}}{\varepsilon} = L_{\omega_0}\varphi,$$

where $L_{\omega_0}\colon C^{k+4,\alpha}(M)_{\mathbb{R}} \to C^{k,\alpha}(M)_{\mathbb{R}}$ is the Lichnerowicz operator for the Kähler metric ω_0 (cf. [23], [14]). Then by (4) the Fréchet derivative $D_\omega\beta_{q,\omega}$ of $\beta_{q,\omega}$ with respect to ω at $(q, \omega) = (0, \omega_0)$ is written as

$$\left\{(D_\omega\beta_{q,\omega})(\sqrt{-1}\,\partial\bar{\partial}\varphi)\right\}_{|(q,\omega)=(0,\omega_0)} = \left\{(D_\omega\sigma_\omega)(\sqrt{-1}\,\partial\bar{\partial}\varphi)\right\}_{|\omega=\omega_0} = L_{\omega_0}\varphi,$$

where L_{ω_0} is an invertible operator by the triviality of H. By setting $q := 1/\xi$, we move q in the half line $\mathbb{R}_{\geq 0} := \{0\} \cup \{1/\xi\,;\, \xi > 0\}$. Replacing q by q^2 if necessary, we apply the implicit function theorem to the map $(q, \omega) \mapsto \beta_{q,\omega}$. (The required regularity for this map is rather delicate: By using [3], Theorem 1.5 and §2.c, we can write both $B_{\xi,\omega}$ and its ω-derivatives as integrals similar to (18) in [50]. Then the estimate of the remainder term in the asymptotic expansion for $B_{m,\omega}$ in [50], Theorem 1, is valid also for the asymptotic expansion of $B_{\xi,\omega}$ and its ω-derivatives. However, for continuity of $\beta_{q,\omega}$ and its ω-derivative, more delicate estimates are necessary. Related to Nash–Moser's process, this will be treated elsewhere.) Then we have openness of the solutions for the one-parameter family of equations

$$\beta_{q,\omega} = 0, \quad q \geq 0, \tag{7}$$

i.e., there exists a one-parameter family of $C^{k+2,\alpha}$ solutions $\omega = \omega(q)$, $0 \leq q < \varepsilon$, for (7) with $\omega(0) = \omega_0$, where ε is sufficiently small. Hence by Fact 3 in Section 5, (M, L^m) is Chow–Mumford stable for all integers $m > 1/\varepsilon$. □

7. The case where M admits symmetries

In this section we consider a polarized manifold (M, L) with an extremal Kähler metric ω_0 in $c_1(L)_{\mathbb{R}}$. Then following Section 6, a result on stability in [32] will be discussed. Let $\mathcal{V}$ be the associated extremal Kähler vector field on M. Assume that H is possibly of positive dimension. Then the identity component K of the isometry group of (M, ω_0) is a maximal compact subgroup in H. Let $\mathfrak{z}$ be the Lie algebra of the identity component Z of the center of K.

Step 1. We assume that the algebraic torus T in Section 5 is the complexification $Z^{\mathbb{C}}$ of Z in H, so that the Lie algebra $\mathfrak{t}$ of T coincides with the Lie algebra $\mathfrak{z}^{\mathbb{C}}$ of $Z^{\mathbb{C}}$. Let $q \in \mathbb{R}_{\geq 0}$, where we set $q = 1/\xi$ for the part $0 < \xi \in \mathbb{R}$. Take an element $\mathcal{W}$ in the Lie algebra $\mathfrak{z}$. Let ω be a Kähler metric in the class $c_1(L)_{\mathbb{R}}$. Then by setting $\mathcal{Y} := iq^2(\mathcal{V} + \mathcal{W})$ for $i := \sqrt{-1}$, we can now consider the following modified weighted Bergman kernel

$$\beta_{q,\omega,\mathcal{W}} := 2\xi\left(1 + \frac{2q}{3}\Delta_\omega\right)(B_{\xi,\omega,\mathcal{Y}} - C_{\xi,\mathcal{Y}}), \quad \xi \in \mathbb{C}^*,$$

where we used (5) and (6) in defining $B_{\xi,\omega,\mathcal{Y}}$ and $C_{\xi,\mathcal{Y}}$. For the Kähler manifold (M, ω), consider the Hamiltonian function $\hat{\sigma}_\omega \in C^\infty(M)_{\mathbb{R}}$ for $\mathcal{V}$ characterized by

$$\mathcal{V} \lrcorner \omega = \bar{\partial}\hat{\sigma}_\omega \quad \text{and} \quad \int_M \sigma_\omega\, \omega^n = \int_M \hat{\sigma}_\omega\, \omega^n.$$

For each $\mathcal{Z} \in \mathfrak{z}$ the associated Hamiltonian function $f_{\mathcal{Z},\omega} \in C^\infty(M)_{\mathbb{R}}$ is uniquely characterized by the identities $\mathcal{Z} \lrcorner \omega = \bar{\partial} f_{\mathcal{Z},\omega}$ and $\int_M f_{\mathcal{Z},\omega}\, \omega^n = 0$. Note that $\mathcal{V} \in \mathfrak{z}$. Then also in this case, as in (4), we obtain

$$\beta_{q,\omega,\mathcal{W}} = \sigma_\omega - \hat{\sigma}_\omega - f_{\mathcal{W},\omega} + O(q). \tag{8}$$

Choose an arbitrary nonnegative integer ℓ with a real number $0 < \alpha < 1$. For the space $\mathcal{F}_\ell$ of all K-invariant functions $f \in C^{\ell,\alpha}(M)_{\mathbb{R}}$ satisfying $\int_M f\omega_0^n = 0$, by setting $\mathcal{N} := \operatorname{Ker} L_{\omega_0} \cap \mathcal{F}_\ell$, we have the identification

$$\theta : \mathfrak{z} \cong \mathcal{N}, \quad \mathcal{Z} \leftrightarrow \theta(\mathcal{Z}) := f_{\mathcal{Z},\omega_0},$$

where $\mathcal{N}$ is independent of the choice of ℓ. Then the vector space $\mathcal{F}_\ell$ is written as a direct sum $\mathcal{N} \oplus \mathcal{N}_\ell^\perp$, where $\mathcal{N}_\ell^\perp$ is the space of all functions f in $\mathcal{F}_\ell$ such that $\int_M f\nu\omega_0^n = 0$ for all $\nu \in \mathcal{N}$. For an arbitrary integer $k \geq 0$, we make a small perturbation of ω_0 by varying ω in the space $\{\omega_0 + \sqrt{-1}\,\partial\bar{\partial}\varphi\,;\ \varphi \in \mathcal{N}_{k+4}^\perp\}$. Since ω_0 is an extremal Kähler metric, we see from (8) that $\beta_{q,\omega,\mathcal{W}}$ vanishes at $(q, \omega, \mathcal{W}) = (0, \omega_0, 0)$, i.e.,

$$\beta_{0,\omega_0,0} = 0 \quad \text{on } M.$$

Again from (8) we see that the Fréchet derivatives $D_\omega \beta_{q,\omega,\mathcal{W}}$ and $D_{\mathcal{W}} \beta_{q,\omega,\mathcal{W}}$ of $\beta_{q,\omega,\mathcal{W}}$ at $(q, \omega, \mathcal{Y}) = (0, \omega_0, 0)$ are

$$\begin{cases} \{(D_\omega \beta_{q,\omega,\mathcal{W}})(\sqrt{-1}\,\partial\bar{\partial}\varphi)\}_{|(q,\omega,\mathcal{W})=(0,\omega_0,0)} = L_{\omega_0}\varphi, & \varphi \in \mathcal{N}_{k+4}^{\perp}, \\ (D_{\mathcal{W}} \beta_{q,\omega,\mathcal{W}})_{|(q,\omega,\mathcal{W})=(0,\omega_0,0)} = -\theta, & \text{on } \mathcal{N}. \end{cases}$$

Since $L_{\omega_0} : \mathcal{N}_{k+4}^{\perp} \to \mathcal{N}_k^{\perp}$ is invertible and since θ is an isomorphism, the implicit function theorem is now applicable to the map: $(q, \omega, \mathcal{W}) \mapsto \beta_{q,\omega,\mathcal{W}}$. Then for some $0 < \varepsilon \ll 1$ we can write

$$\omega = \omega(q) \quad \text{and} \quad \mathcal{W} = \mathcal{W}(q), \quad 0 \le q \le \varepsilon,$$

solving the one-parameter family of equations

$$\beta_{q,\omega,\mathcal{W}} = 0, \quad q \ge 0,$$

in $(\omega, \mathcal{W})$. Hence by setting $\mathcal{Y}(q) := iq^2(\mathcal{V} + \mathcal{W}(q))$ for $q = 1/m$, the m-th weighted Bergman kernel $B_{m,\omega(q),\mathcal{Y}(q)}$ is constant on M for all $m > 1/\varepsilon$. Then by Fact 4 in Section 5, (M, L^m) is Chow–Mumford T-stable for all $m > 1/\varepsilon$.

Step 2. Though we assumed that T coincides with $Z^{\mathbb{C}}$ in the first step, it is better to choose T as small as possible. Then for a sufficiently small positive real number ε, the algebraic torus T in $Z^{\mathbb{C}}$ generated by

$$\bigcup_{1/\varepsilon < m \in \mathbb{Z}} \{\mathcal{V} + \mathcal{W}(q), i\mathcal{V} + i\mathcal{W}(q)\}$$

is a good choice, where $q = 1/m$ and $i := \sqrt{-1}$.

8. Concluding remarks

For the conjecture in Section 3 the "if" part is known to be a difficult problem and it is of particular interest. Assuming that a polarized manifold (M, L) is asymptotically Chow–Mumford stable, we are asked whether there exist Kähler metrics of constant scalar curvature in $c_1(L)_{\mathbb{R}}$. For simplicity, we consider the case where H is trivial. Then by Fact 3 the equation

$$\beta_{q,\omega} = 0 \tag{9}$$

has solutions $(q, \omega) = (1/m, \omega_m)$, $m \gg 1$, while each ω_m is an m-th balanced metric for (M, L). Moreover, the triviality of H implies that ω_m is the only m-th balanced metric for (M, L). Then we are led to study the graph

$$\mathcal{E} := \{(q, \omega)\,;\ \beta_{q,\omega} = 0\}$$

in $\mathbb{C} \times \mathcal{K}$, where $\mathcal{K}$ is the space of all Kähler metrics in the class $c_1(L)_{\mathbb{R}}$. Let $\mathcal{H}$, $\mathcal{M}$ be the sets of all Hermitian metrics for L, V_m, respectively. Consider the Fréchet derivative $D_\omega \beta_{q,\omega}$ at $(q, \omega) = (1/m, \omega_m) \in \mathcal{E}$, where $m \gg 1$. With the same setting of differentiability as in Section 6, the Fréchet derivative will be shown to be invertible.

Let us give a rough idea how the invertibility can be proved. It suffices to show that the operator $D_\omega B_{m,\omega}$ is invertible at $\omega = \omega_m$. Choose a Hermitian metric h_m for L such that $c_1(L; h_m) = \omega_m$. Put $\Psi := \{\psi \in C^\infty(M)_{\mathbb{R}}\,;\, \int_M \psi \omega_m^n = 0\}$. Let $\varphi \in \Psi$. Then the $\mathbb{R}$-orbit in $\mathcal{H}$ through h_m written in the form

$$h_{\varphi;t} := e^{-t\varphi} h_m, \quad t \in \mathbb{R}, \tag{10}$$

projects to the $\mathbb{R}$-orbit $\omega_{\varphi,t} := \omega_m + \sqrt{-1}\, t\, \partial\bar{\partial}\varphi$, $t \in \mathbb{R}$, in $\mathcal{K}$ through ω_m. Note also that every Hermitian metric h for L induces a Hermitian pairing $\langle\ ,\ \rangle_{L^2}$ which will be denoted by $(V_m, \tilde{h})$ (cf. Section 4). Choose an orthonormal basis $\{\sigma_1, \sigma_2, \ldots, \sigma_{N_m}\}$ for $(V_m, \tilde{h}_m)$. Let Φ be the space of all $\varphi \in \Psi$ such that

$$(d/dt)(\tilde{h}_{\varphi,t})_{|t=0} = 0 \quad \text{on } V_m.$$

In other words, if $\varphi \in \Phi$ then the basis $\{\sigma_1, \sigma_2, \ldots, \sigma_{N_m}\}$ infinitesimally remains to be an orthonormal basis for $(V_m, \tilde{h}_{\varphi,t})$ at $t = 0$ with t perturbed a little. Since

$$B_{m,\omega} := |\sigma_1|_h^2 + |\sigma_2|_h^2 + \cdots + |\sigma_{N_m}|_h^2$$

is obtained from the contraction of $\Sigma := \sigma_1\bar{\sigma}_1 + \sigma_2\bar{\sigma}_2 + \cdots + \sigma_{N_m}\bar{\sigma}_{N_m}$ by h^m, and since Σ is fixed by the infinitesimal action $(d/dt)(h_{\varphi,t})$ at $t = 0$ in (10), we obtain

$$(d/dt)(B_{m,\omega_{\varphi,t}})_{|t=0} = (d/dt)(h^m_{\varphi,t})_{|t=0} = -m\varphi. \tag{11}$$

Let $\mathbb{H}$ denote the set of all Hermitian metrics on the vector space V_m. Then we have a natural projection $\pi : \Psi \to T_{\tilde{h}_m}\mathbb{H}$ defined by

$$(d/dt)(h_{\varphi,t})_{|t=0} \mapsto (d/dt)(\tilde{h}_{\varphi,t})_{|t=0}.$$

In view of $\dim \mathbb{H} < +\infty$, it is now easy to see that this map is surjective. Since the kernel of this map is Φ, we obtain

$$\Psi/\Phi \cong T_{\tilde{h}_m}\mathbb{H}. \tag{12}$$

Finally by (11) and (12), the uniqueness of the m-th balanced metric (by $H = \{1\}$) implies the required invertibility of the Fréchet derivative (see also Donaldson [16]). □

Now we can apply the implicit function theorem to obtain an open neighborhood U of $1/m$ in $\mathbb{C}$ such that, for each $\xi \in U$, there exists a unique Kähler metric $\omega(\xi)$ in $c_1(L)_{\mathbb{R}}$ satisfying $\beta_{\xi,\omega(\xi)} = 0$.

Assuming non-existence of Kähler metrics of constant scalar curvature, we have some possibility that, by an argument as in Nadel, destabilizing objects are obtained by studying the behavior of the solutions along the boundary point of $\mathcal{E}$.

Finally, there are many interesting topics, which are related to ours, such as the geometry of Kähler potentials [10], [9], [38], [40] (see also [28]). The uniqueness of extremal Kähler metrics modulo the action of holomorphic automorphisms in compact cases is recently given in [12] (cf. [2], [14]). New obstructions to semistability of manifolds or to the existence of extremal metrics are done by [19], [29], [12]. The concept of K-stability, introduced by Tian [45] and reformulated by Donaldson [15], is deeply related to the Hilbert–Mumford stability criterion, and various kinds of works are actively done related to algebraic geometry.

References

[1] Aubin, T., *Non-linear Analysis on Manifolds*. Grundlehren Math. Wiss. 252, Springer-Verlag, New York 1982.

[2] Bando, S., Mabuchi, T., Uniqueness of Einstein Kähler metrics modulo connected group actions. In *Algebraic Geometry* (Sendai, 1985) Adv. Stud. Pure Math. 10, Kinokuniya and North-Holland, Tokyo and Amsterdam 1987, 11–40.

[3] Boutet de Monvel, L., Sjöstrand, J., Sur la singularité des noyaux de Bergman et de Szegö. In *Équations aux Dérivées Partielles de Rennes 1975*, *Astérisque* **34-35** (1976), 123–164.

[4] Catlin, D., The Bergman kernel and a theorem of Tian. In *Analysis and Geometry in Several Complex Variables* (ed. by G. Komatsu, M. Kuranishi), Trends in Math., Birkhäuser, Boston 1999, 1–23.

[5] Calabi, E., The space of Kähler metrics. In *Proceedings of the International Congress of Mathematicians* (Amsterdam, 1954), Vol. II, Erven P. Noordhoff N.V. and North-Holland Publishing Co., Groningen and Amsterdam 1956, 206–207.

[6] Calabi, E., On Kähler manifolds with vanishing canonical class. In *Algebraic Geometry and Topology* (A symposium in honor of S. Lefschety), Princeton University Press, Princeton, NJ, 1955, 78–89.

[7] Calabi, E., Improper affine hyperspheres of convex type and a generalization of a theorem by K. Jörgens. *Michigan Math. J.* **5** (1958), 105–126.

[8] Calabi, E., Extremal Kähler metrics II. In *Differential Geometry and Complex Analysis* (ed. by I. Chavel, H. M. Farkas), Springer-Verlag, Berlin 1985, 95–114.

[9] Calabi, E., Chen X., The space of Kähler metrics, II. *J. Differential Geom.* **61** (2002), 173–193.

[10] Chen, X., The space of Kähler metrics. *J. Differential Geom.* **56** (2000), 189–234.

[11] Chen, X., Tian G., Ricci flow on Kähler-Einstein surfaces. *Invent. Math.* **147** (2002), 487–544.

[12] Chen, X., Tian G., Uniqueness of extremal Kähler metrics. *C. R. Acad. Sci. Paris* **340** (2005), 287–290.

[13] Donaldson, S. K., Infinite determinants, stable bundles and curvature. *Duke Math. J.* **54** (1987), 231–247.

[14] Donaldson, S. K., Scalar curvature and projective embeddings, I. *J. Differential Geom.* **59** (2001), 479–522.

[15] Donaldson, S. K., Scalar curvature and stability of toric varieties. *J. Differential Geom.* **62** (2002), 289–349.

[16] Donaldson, S. K., Scalar curvature and projective embeddings, II. *Quart. J. Math.* **56** (2005), 345–356.

[17] Fogarty, J., Truncated Hilbert functors. *J. Reine und Angew. Math.* **234** (1969) 65–88.

[18] Futaki, A., An obstruction to the existence of Einstein Kähler metrics. *Invent. Math.* **73** (1983), 437–443.

[19] Futaki, A., Asymptotic Chow semi-stability and integral invariants. *Internat. J. Math.* **15** (2004), 967–979.

[20] Gieseker, D., Global moduli for surfaces of general type. *Invent. Math.* **43** (1977), 233–282.

[21] Hirachi, K., Zelditch expansion. A private note on the homepage http://www.ms.u-tokyo.ac.jp/ hirachi/papers, 1999.

[22] Kobayashi, S., *Differential Geometry of Complex Vector Bundles.* Publ. Math. Soc. Japan 14, Iwanami and Princeton University Press, Tokyo and Princeton 1987.

[23] Lichnerowicz, A., Sur les transformations analytiques des variétés kählériennes. *C. R. Acad. Sci. Paris* **244** (1957), 3011–3014.

[24] Lu, Z., On the lower order terms of the asymptotic expansion of Tian-Yau-Zelditch *Amer. J. Math.* **122** (2000), 235–273.

[25] Lübke, M., Stability of Einstein-Hermitian vector bundles. *Manuscripta Math.* **42** (1983), 245–257.

[26] Luo, H., Geometric criterion for Gieseker-Mumford stability of polarized manifolds. *J. Differential Geom.* **49** (1998), 577–599.

[27] Lübke, M., Teleman, T., *The Kobayashi-Hitchin Correspondence*. World Scientific, Singapore 1995.

[28] Mabuchi, T., Some symplectic geometry on compact Kähler manifolds (I). *Osaka J. Math.* **24** (1987), 227–252.

[29] Mabuchi, T., An obstruction to asymptotic semistability and approximate critical metrics. *Osaka J. Math.* **41** (2004), 463–472.

[30] Mabuchi, T., Stability of extremal Kähler manifolds. *Osaka J. Math.* **41** (2004), 563–582.

[31] Mabuchi, T., An energy-theoretic approach to the Hitchin-Kobayashi correspondence for manifolds, I. *Invent. Math.* **159** (2005), 225–243.

[32] Mabuchi, T., An energy-theoretic approach to the Hitchin-Kobayashi correspondence for manifolds, II. arxiv: math.DG/0410239.

[33] Mabuchi, T., The Chow-stability and Hilbert-stability in Mumford's geometric invariant theory. In preparation.

[34] Matsushima, Y., Sur la structure du groupe d'homéomorphismes analytiques d'une certaine variété kählérienne. *Nagoya Math. J.* **11** (1957), 145–150.

[35] Mumford, D., Stability of projective varieties. *L'Enseignment Math.* **23** (1977), 39–110.

[36] Mumford, D., Fogarty, J., Kirwan, F., *Geometric Invariant Theory*. 3rd enlarged ed., Ergeb. Math. Grenzgeb. 34, Springer-Verlag, Berlin 1994.

[37] Nadel, A. M., Multiplier ideal sheaves and Kähler-Einstein metrics of positive scalar curvature. *Ann. of Math.* **132** (1990), 549–596.

[38] Phong, D. H., Sturm, J., Stability, energy functional, and Kähler-Einstein metrics. *Comm. Anal. Geom.* **11** (2003), 565–597.

[39] Phong, D. H., Sturm, J., Scalar curvature, moment maps, and the Deligne pairing. *Amer. J. Math.* **126** (2004), 693–712.

[40] Semmes, S., Complex Monge-Ampère and symplectic manifolds. *Amer. J. Math.* **114** (1992), 495–550.

[41] Siu, Y. T., The existence of Kähler-Einstein metrics on manifolds with anticanonical line bundle and a suitable finite symmetry group. *Ann. of Math.* **127** (1988), 585–627.

[42] Tian, G., On Kähler-Einstein metrics on certain Kähler manifolds with $C_1(M) > 0$. *Invent. Math.* **89** (1987), 225–246.

[43] Tian, G., On a set of polarized Kähler metrics on algebraic manifolds. *J. Differential Geom.* **32** (1990), 99–130.

[44] Tian, G., On Calabi's conjecture for complex surfaces with positive first Chern class. *Invent. Math.* **101** (1990), 101–172.

[45] Tian, G., Kähler-Einstein metrics with positive scalar curvature. *Invent. Math.* **130** (1997), 1–37.

[46] Tian, G., Yau S.-T., Kähler-Einstein metrics on complex surfaces with $C_1 > 0$. *Comm. Math. Phys.* **112** (1987), 175–203.

[47] Uhlenbeck, K., Yau S.-T., On the existence of Hermitian-Yang-Mills connections in stable vector bundles. *Comm. Pure Appl. Math.* **39** (1986), S257–S293.

[48] Viehweg, E., *Quasi-projective moduli for polarized manifolds*. Ergeb. Math. Grenzgeb. (3) 30, Springer-Verlag, Berlin 1995.

[49] Yau S.-T., On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampère equation I. *Comm. Pure Appl. Math.* **31** (1978), 339–411.

[50] Zelditch, S., Szegö kernels and a theorem of Tian. *Internat. Math. Res. Notices.* **6** (1998), 317–331.

[51] Zhang, S., Heights and reductions of semi-stable varieties. *Compositio Math.* **104** (1996), 77–105.

[52] Wang, X.-J., Zhu, X., Kähler-Ricci solitons on toric manifolds with positive first Chern class. *Adv. Math.* **188** (2004), 87–103.

Department of Mathematics, Osaka University, Toyonaka, Osaka, 560-0043 Japan
E-mail: mabuchi@math.sci.osaka-u.ac.jp

Tropical geometry and its applications

Grigory Mikhalkin*

Abstract. From a formal perspective tropical geometry can be viewed as a branch of geometry manipulating with certain piecewise-linear objects that take over the rôle of classical algebraic varieties. This talk outlines some basic notions of this area and surveys some of its applications for the problems in classical (real and complex) geometry.

Mathematics Subject Classification (2000). 14A99, 14H50, 14N10, 52B20.

Keywords. Tropical geometry, amoebas, patchworking, enumerative geometry.

1. Introduction

From a geometric point of view tropical geometry describes worst possible degenerations of the complex structure on an n-fold X. Such degenerations cause X to collapse onto an n-dimensional (over $\mathbb{R}$, i.e. $\left(\frac{\dim X}{2}\right)$-dimensional) base B which is a piecewise-linear polyhedral complex, see e.g. [13] for a conjectural picture in the special case of Calabi–Yau n-folds.

According to an idea of Kontsevich such degenerations can be useful, in particular, for computations of the Gromov–Witten invariants of X as holomorphic curves degenerate to graphs $\Gamma \subset B$. A similar picture appeared in the work of Fukaya (see e.g. [5]) where graphs come as degenerations of holomorphic membranes.

Tropical geometry formalizes the base B as an ambient variety so that the graphs Γ become curves in B. Some problems in complex and real geometry then may be reduced to problems of tropical geometry which are often much easier to solve, thanks to the piecewise-linear nature of the subject. Local considerations in tropical geometry correspond to some standard models in classical geometry while the combinatorial tropical structure encodes the way to glue these models together. In this sense it may be viewed as an extension of the Viro patchworking [32].

This talk takes a geometric point of view on tropical geometry and surveys its basic notions as well as some of its applications to classical algebro-geometrical problems. The author is indebted to Ya. Eliashberg, M. Kontsevich, A. Losev, B. Sturmfels and O. Viro for many useful conversations on tropical geometry.

*The author is grateful to the Institut Henri Poincaré (Paris) and the IHES for hospitality during the preparation of this talk. The author's research is supported in part by NSERC.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. Tropical algebra

2.1. Tropical vs. classical arithmetics. The term *tropical semirings* was reputedly invented by a group of French computer scientists to commemorate their Brazilian colleague Imre Simon. For our purposes we use just one of these semirings, the tropical semifield $\mathbb{T} = \mathbb{R} \cup \{-\infty\}$ equipped with the operations of addition and multiplication defined by

$$\text{“}a + b\text{”} = \max\{a, b\}, \quad \text{“}ab\text{”} = a + b.$$

We use the quotation marks to distinguish the tropical operations from the classical addition and multiplication on $\mathbb{R}\cup\{-\infty\}$. It is easy to check that the tropical operations are commutative, associative and satisfy the distribution law.

There is no tropical subtraction as the operation “ $+$ ” is idempotent, but we have the tropical division $\text{“}\frac{a}{b}\text{”} = a - b$, $b \neq -\infty$. This makes $\mathbb{T}$ an idempotent semifield. Its additive zero is $0_{\mathbb{T}} = -\infty$, its multiplicative unit is $1_{\mathbb{T}} = 0$.

Remark 2.1. A classical example of a semifield is the semifield $\mathbb{R}_{\geq 0}$ of nonnegative numbers with classical addition and multiplication. We have the map $\log_t : \mathbb{R}_{\geq 0} \to \mathbb{T}$ between semifields $\mathbb{R}_{\geq 0}$ for any $t > 1$. The larger t, the closer is the map $\log_t$ to a homomorphism. More specifically, $\log_t$ is an isomorphism up to the error of $\log_t(2)$ while $\lim_{t\to+\infty} \log_t 2 = 0$. Indeed, we have $\log_t(ab) = \text{“}\log_t(a)\log_t(b)\text{”}$ and

$$\text{“}\log_t(a) + \log_t(b)\text{”} \leq \log_t(a + b) \leq \text{“}\log_t(a) + \log_t(b)\text{”} + \log_t 2$$

by elementary considerations. Thus $\mathbb{T}$ can be considered as the $t \to +\infty$ limit semifield of a family of (mutually isomorphic) semifields $\mathbb{R} \cup \{-\infty\}$ equipped with arithmetic operations induced from $\mathbb{R}_{\geq 0}$ by $\log_t$. In this sense $\mathbb{T}$ is considered to be the result of *dequantization* of classical arithmetics by Maslov et al., see e.g. [15] or [33] for a geometric version of this dequantization.

Remark 2.2. The semifield $\mathbb{T}$ is closely related to non-Archimedean fields K with (real) valuations $\mathrm{val} : K \to \mathbb{T}$. Indeed, val is a valuation if for any $z, w \in K$ we have $\mathrm{val}(z + w) \leq \text{“}\mathrm{val}(z) + \mathrm{val}(w)\text{”}$, $\mathrm{val}(zw) = \text{“}\mathrm{val}(z)\,\mathrm{val}(w)\text{”}$ and $\mathrm{val}^{-1}(-\infty) = 0$. In this sense val is a sub-homomorphism. This non-Archimedean point of view on tropical geometry is taken e.g. in [4], [25], [29] and [30].

2.2. Polynomials, regular and rational functions. There is no subtraction in the semifield $\mathbb{T}$, but we do not need it to form polynomials. A tropical polynomial is a tropical sum of monomials. For a polynomial f in n variables we get $f : \mathbb{T}^n \to \mathbb{R}$,

$$f(x) = \text{“}\sum_j a_j x^j\text{”} = \max(a_j + \langle j, x\rangle), \tag{1}$$

where $x = (x_1, \ldots, x_n) \in \mathbb{T}^n$, $j = (j_1, \ldots, j_n) \in \mathbb{Z}^n$, $x^j = x_1^{j_1} \ldots x_n^{j_n}$, $\langle j, x\rangle = j_1x_1 \ldots j_nx_n$ and $a_j \in \mathbb{T}$.

Remark 2.3. Not all monomials in f are essential. It may happen that for some element j' we have $a_{j'} + \langle j', x\rangle \le a_j + \langle j, x\rangle$ for any $j \ne j'$. Then for any $b_{j'} \le a_j$ we have "$\sum_j a_j x^j$" = "$b_{j'} + \sum_{j\ne j'} a_j x^j$" = "$\sum_{j\ne j'} a_j x^j$". Thus the presentation of a function f as a tropical polynomial is not, in general, unique. Nevertheless, it is very close to being unique.

Note that (1) defines the Legendre transform of a (partially-defined) function on $\mathbb{R}^n$, $-a\colon j \mapsto -a_j$. The function $-a$ is defined only on a finite set, namely the set J which consists of the powers of the monomials in f. We can use involutivity of the Legendre transform to recover the coefficients a_j from the function f. Take the Legendre transform L_f of f and set $\bar a_j = -L_f(j)$, $j \in \mathbb{Z}^n$. It is easy to see that if f were a tropical polynomial then $a_j = -\infty = 0_{\mathbb{T}}$ for all but finitely many $j \in \mathbb{Z}^n$. Furthermore, if the function $-a$ is convex (i.e. if it is a restriction of a convex function on $\mathbb{R}^n$) then $\bar a_j = a_j$. Thus every function $f\colon \mathbb{T}^n \to \mathbb{T}$ obtained from a tropical polynomial has a "maximal" presentation as a tropical polynomial (clearly, $\bar a_j \ge a_j$).

Once we have polynomials we may form rational functions as the tropical quotients, i.e. the differences of two polynomials. The tropical quotient of two monomials is called a $\mathbb{Z}$-*affine* function. Clearly a $\mathbb{Z}$-affine function is an affine-linear function $\mathbb{R}^n \to \mathbb{R}$ with an integer slope.

A rational function $h = "\frac{f}{g}"$ is defined for every $x \in \mathbb{T}^n$ such that $g(x) \ne -\infty$. At such x h is called *finite* (more precisely, h is called finite at $x \in \mathbb{T}^n$ if it *can be presented* as "$\frac{f}{g}$" with $g(x) \ne -\infty$). A rational function h is called *regular* at $x \in \mathbb{T}^n$ if it is finite at x and there exist a $\mathbb{Z}$-affine function ϕ finite at x, an open neighborhood $U \ni x$ and a tropical polynomial $p\colon \mathbb{T}^n \to \mathbb{T}$ such that $h(y) = p(y) + \phi(y)$ for all $y \in U$. (Here we consider the Euclidean topology on $\mathbb{T}^n = [-\infty, +\infty)^n$.)

All functions $U \to \mathbb{T}$ that are restrictions of rational functions on $\mathbb{T}^n$ that are regular at every point of U form a semiring $\mathcal{O}(U)$. Constant functions give a natural embedding $\mathbb{T} \subset \mathcal{O}(U)$ and thus $\mathcal{O}(U)$ is a *tropical algebra*. In this way we get a sheaf $\mathcal{O}$ of tropical algebras on $\mathbb{T}^n$ called *the structure sheaf*.

3. Geometry: tropical varieties

3.1. Hypersurfaces. To every tropical polynomial f one may associate its hypersurface $V_f \subset \mathbb{T}^n$ that, by definition, consists of all points x where "$\frac{1_{\mathbb{T}}}{f}$" is not regular. This is a piecewise-linear object in $\mathbb{T}^n$.

Example 3.1. Figures 1 and 2 depict some hypersurfaces in $\mathbb{T}^2$ (i.e. planar curves).

The left-hand side of Figure 1 is given by the polynomial "$1 + 0x + 0y$". It is a line in the tropical plane $\mathbb{T}^2$. The right-hand side of Figure 1 is given by the polynomial "$10 + 5.5x + 0x^2 + 8.5y + 6.5y^2 + 4.5xy$", it is a conic. The line and the conic here intersect at two distinct points.

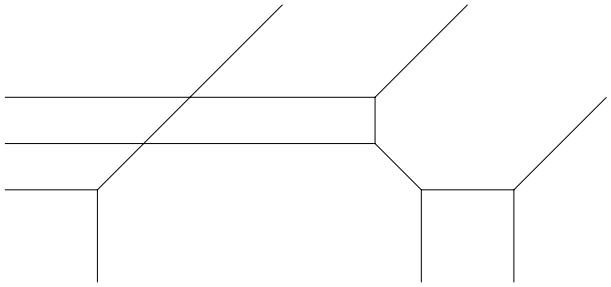


Figure 1. Tropical line and conic.

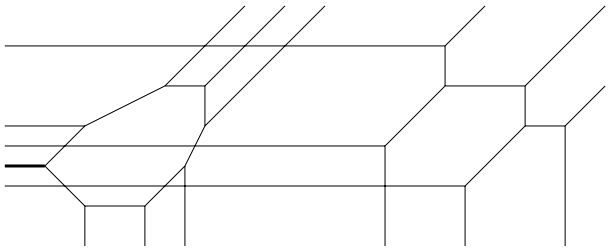


Figure 2. Two tropical cubic curves.

The left-hand side of Figure 2 is given by the polynomial

$$\text{“}5 + 4x + 2.25x^2 + 0x^3 + 4y + 2.5xy + 1x^2y + 3y^2 + 1.5xy^2 + 1.5y^3\text{”}. \quad (2)$$

The right-hand side is given by the polynomial

$$\text{“}17.5 + 12.25x + 7x^2 + 0x^3 + 16.75y + 12xy + 5.5x^2y + 15.5y^2 + 10xy^2 + 13y^3\text{”}.$$

It is easy to see that a tropical hypersurface V_f is a union of convex $(n-1)$-dimensional polyhedra called the facets of V_f. Each facet has *integer slope*, i.e. is parallel to a hyperplane in $\mathbb{R}^n$ defined over $\mathbb{Z}$. For every facet P there exist two monomials $a_{j_1}x^{j_1}$ and $a_{j_2}x^{j_2}$ of f that are equal along P and such that they are greater than any other monomial of f in the (relative) interior of P. The greatest common divisor of the components of the vector $j_2 - j_1$ is called *the weight* of P. E.g. the horizontal edge in Figure 2 adjacent to the leftmost vertex has weight 2 while all other edges have weight 1.

The intersection of any pair of facets is the face of both facets. Furthermore, at every $(n-2)$-dimensional face Q we have the following *balancing property*.

Property 3.2. *Let $P_1, \dots, P_l$ be the facets adjacent to Q and let v_j be the integer covectors whose kernels are parallel to P_j and whose gcd is equal to the weight of P_j.*

In addition we require that the orientation of $v_1, \dots, v_l$ is consistent with a choice of direction around Q. Then we have

$$\sum_{j=1}^{l} v_j = 0.$$

3.2. Integer polyhedral complexes. This balancing property may also be generalized to some piecewise-linear polyhedral complexes X of arbitrary codimension in $\mathbb{T}^n$. A *integer convex polyhedron* in $\mathbb{T}^n$ is the set defined by a finite number of inequalities of the type

$$\langle j, x\rangle \le c,$$

where $x \in \mathbb{T}^n$, $j \in \mathbb{Z}^n$ and $c \in \mathbb{R}$. Here the expression $\langle j, x\rangle$ stands for the scalar product. It may happen that $\langle j, x\rangle$ is an indeterminacy (for some $x \in \mathbb{T} \setminus \mathbb{T}^\times$), we include such x into the polygon. Equivalently, an integer convex polyhedron in $\mathbb{T}^n$ is the closure of a convex polyhedron (bounded or unbounded) in $\mathbb{R}^n$ such that the slopes of all its faces (including the polyhedron itself) are integers. The dimension of an integer convex polyhedron is its topological dimension.

An integer piecewise-linear polyhedral complex X of dimension k in $\mathbb{T}^n$ is the union of a finite collection of integer convex polyhedra of dimension k called *the facets of X* such that the intersection $\bigcap_{j=1}^{l} P_j$ of any finite number of facets is the common face of P_j. We may equip the facets of X with natural numbers called the weights.

The complex X is called balanced if the following property holds.

Property 3.3. *Let Q be a face of dimension $k-1$ and $P_1, \dots, P_l$ be the facets adjacent to Q. The affine-linear space containing Q defines a linear projection $\lambda\colon \mathbb{R}^n \to \mathbb{R}^{n-k+1}$. The image $q = \lambda(Q)$ is a point while the images of $p_j = \lambda(P_j)$ are intervals in $\mathbb{R}^{n-k+1}$ adjacent to q. Let $v_j \in \mathbb{Z}^n$ be the primitive integer vector parallel to p_j in the direction outgoing from q multiplied by the weight of P_j. We have*

$$\sum_{j=1}^{l} v_j = 0.$$

It is easy to see that if $k = n-1$ then Properties 3.2 and 3.3 are equivalent.

3.3. Contractions. Let $f\colon \mathbb{T}^n \to \mathbb{T}$ be a polynomial. We define its *full graph*

$$\Gamma_f \subset \mathbb{T}^n \times \mathbb{T}$$

to be the hypersurface defined by "$y + f(x)$". Note that Γ_f can be obtained from the set-theoretical graph of f by attaching the intervals $[(x, -\infty), (x, f(x))]$ for all x from the hypersurface V_f (i.e. those x where "$\frac{0}{f}$" is not regular), cf. Figure 3 for the full graphs of "$x + 0$" and "$x^2 + x + 1$".

Thus, unlike the classical situation, the full graph of a map is different from the domain of the map. We define the *principal contraction*

$$\delta_f : \Gamma_f \to \mathbb{T}^n \tag{3}$$

associated to f to be the projection onto $\mathbb{T}^n$.

To get a general contraction one iterates this procedure. Suppose that a contraction $\gamma : V \to \mathbb{T}^n$ is already defined. We have $V \subset \mathbb{T}^N$.

If $g : V \to \mathbb{T}$ is a regular function (in N variables) then we can define the full graph $\Gamma_g \subset V \times \mathbb{T}$ as the union of the set-theoretical graph of g with all intervals $[(x, -\infty), (x, g(x))]$ such that "$\frac{0}{g}$" is not regular at x (i.e. at a neighborhood of x in V). The map $\delta_g : \Gamma_g \to V$ is the projection onto V.

The map δ_g is called a principal contraction to V. A general contraction is a composition of principal contractions.

Furthermore, one may associate *the weights* to the new facets of Γ_g by setting them equal to the *orders* of the pole of "$\frac{0}{g}$". Here we say that the order of a pole of a rational function is at least n at x if it can be locally presented as a tropical product of n rational functions that are not regular at x. Then we define the order of the pole to be the largest n with this property.

Definition 3.4. The *order of zero* of g at x is the order of the pole of "$\frac{0}{g}$" at x.

This definition is consistent with the definition of tropical hypersurfaces.

The facets of Γ_g contained in the set-theoretical graph of g inherit their weights from the weights of the corresponding facets of V.

Remark 3.5. Contractions may be used to define counterparts of Zariski open sets in tropical geometry. We define the complements of tautological embeddings $\mathbb{T}^k \to \mathbb{T}^n$, $k \leq n$, to be Zariski open.

To define the principal open set corresponding to a polynomial f we consider the contraction $\delta_f : \Gamma_f \to \mathbb{T}^n$ and take

$$D_f = \Gamma_f \cap (\mathbb{T}^n \times \mathbb{T}^\times).$$

This together with the map $\delta_f|_{D_f}$ is the principal open set associated to f.

Example 3.6. The principal open set associated to "$x + a$" $D_{"x+a"} \to \mathbb{T}$ of "$x + a$", $a \in \mathbb{T}^\times$, can be interpreted as the result of puncturing $\mathbb{T}$ at a finite point $a \subset \mathbb{T}$. Figure 3 depicts $D_{"x+0"}$ and $D_{"x^2+x+1"}$. The corresponding maps are projections onto the x-axis.

Proposition 3.7. *If $\delta : \mathbb{T}^N \supset V \to \mathbb{T}^n$ is a contraction then $V \subset \mathbb{T}^N$ satisfies Property* 3.3.

Remark 3.8. In fact Proposition 3.7 can be used to define the weights of the facets of V and thus the orders of the zeroes of polynomials on V.

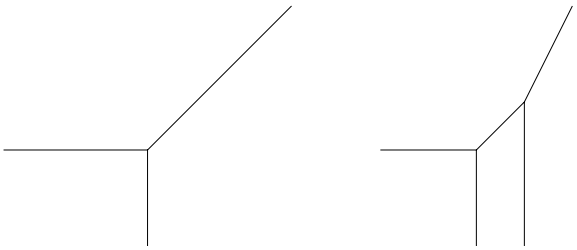


Figure 3. Once and twice punctured affine lines $\mathbb{T}$.

3.4. Tropical varieties and tropical morphisms. A map $\mathbb{R}^N \to \mathbb{R}^M$ is called *integer affine-linear* if it is a composition of a linear map defined over $\mathbb{Z}$ and a translation by an arbitrary vector in $\mathbb{R}^M$. Such a map can be extended to a partially defined map $\mathbb{T}^N \to \mathbb{T}^M$ by taking the closure. Note that integer affine-linear maps leave the class of integer piecewise-linear polyhedral complexes invariant. Furthermore, such maps take facets to facets (at least for some presentation of the image as an integer piecewise-linear polyhedral complex) and respect Property 3.3.

Let X be a topological space together with an atlas $\{U_\alpha\}$, $\phi_\alpha \colon U_\alpha \to \mathbb{T}^n$, $U_\alpha \subset X$.

Definition 3.9. We say that X is a *tropical variety* of dimension n if the following conditions hold.

1. Each ϕ_α is a contraction to an open subset of $\mathbb{T}^n$. More precisely, there is a contraction $\delta_\alpha \colon V_\alpha \to \mathbb{T}^n$, $V_\alpha \subset \mathbb{T}^{N_\alpha}$ and an open embedding $\Phi_\alpha \colon U_\alpha \to V_\alpha$ such that $\phi_\alpha = \delta_\alpha \circ \Phi_\alpha$.
2. The overlapping maps $\Phi_\beta \circ \Phi_\alpha^{-1}$ are induced by the integer affine-linear maps $\mathbb{R}^{N_\alpha} \to \mathbb{R}^{N_\beta}$.
3. For every point $x \in X$ in the interior of a facet of X there exists a chart ϕ_α such that $x \in U_\alpha$ and ϕ_α embeds some neighborhood of x into $\mathbb{T}^n$.
4. There exist a finite covering of X by open sets $W_j \subset X$ such that for every j there exists α such that $W_j \subset U_\alpha$ and the closure of $\Phi_\alpha(W_j)$ in $\mathbb{T}^{N_\alpha}$ is contained in $\Phi_\alpha(U_\alpha)$.

The last condition ensures *completeness* of the tropical structure on X.

A function on a subset W of X is called regular at a point $x \in W$ if it is locally a pull-back of a regular function in a neighborhood of $\Phi_\alpha(x) \in \mathbb{T}^{N_\alpha}$. In this way we get the structure sheaf $\mathcal{O}_X$ of regular functions on X.

A point $x \in X$ is called *finite* if it is mapped to $\mathbb{R}^{N_\alpha} \subset \mathbb{T}^{N_\alpha}$. Note that finiteness does not depend on the choice of a chart. At finite points of x we have the notion of an *integer tangent* vector to X. This comes from tangent vectors to $\Phi_\alpha(U_\alpha)$ at $\Phi_\alpha(x) \in \mathbb{R}^{N_\alpha}$ after identification with the corresponding counterparts for different charts $U_\beta \ni z$ under the differential of the overlapping maps.

Example 3.10. The space $\mathbb{T}^n$ is a n-dimensional tropical variety tautologically. The *projective n-space* $\mathbb{TP}^n$ is defined as the quotient of $\mathbb{T}^{n+1} \setminus (-\infty, \dots, -\infty)$ by the equivalence relation $(x_0, \dots, x_n) \sim (\text{“}\lambda x_0\text{”}, \dots, \text{“}\lambda x_n\text{”}) = (\lambda + x_0, \dots, \lambda + x_n)$, where $\lambda \in \mathbb{T}^\times$.

The space $\mathbb{TP}^n$ is a tropical variety since it admits (as in the classical case) $n+1$ affine charts to $\mathbb{T}^n$ by dividing all coordinates by x_j as long as $x_j \neq -\infty$. This is an example of a compact tropical variety. There is a well-defined notion of a hypersurface of degree d in $\mathbb{TP}^n$. It is given by a homogeneous polynomial of degree d in $n+1$ variables. This polynomial can be translated to an ordinary polynomial in every affine chart of $\mathbb{TP}^n$.

In a similar way, one may construct tropicalizations of other toric varieties. The finite part of all tropical toric varieties is $(\mathbb{T}^\times)^n \approx \mathbb{R}^n$.

Proposition 3.11. *If X and Y are tropical varieties then $X \times Y$ has a natural structure of tropical variety of dimension* $\dim X + \dim Y$.

Definition 3.12. Let $f\colon X \to Y$ be a map between tropical varieties (not necessarily of the same dimension). We say that f is a *linear tropical morphism* if for every $x \in X$ there exist charts $U^X_\alpha \ni x$ and $U^Y_\beta \ni f(x)$ such that $\Phi^Y_\beta \circ f \circ (\Phi^X_\alpha)^{-1}$ is induced by an integer affine-linear map $\mathbb{R}^{N_\alpha} \to \mathbb{R}^{N_\beta}$.

The map f is called a *regular tropical morphism* if $\Phi^Y_\beta \circ f \circ (\Phi^X_\alpha)^{-1}$ is given by N_β rational functions on $\mathbb{T}^{N_\alpha}$ that are regular on $\Phi^X_\alpha(U^X_\alpha)$.

Clearly any linear morphism is a regular morphism. A regular morphism $f\colon X \to Y$ defines a map $\mathcal{O}_Y(U) \to \mathcal{O}_X(f^{-1}(U))$. This map can be interpreted as a homomorphism of tropical algebras (defined over the semifield $\mathbb{T}$).

Proposition 3.13. *If $f\colon X \to Y$ is a linear tropical morphism then its (set-theoretical) graph is a* $(\dim X)$*-dimensional tropical variety.*

3.5. Equivalence of tropical varieties. Different tropical varieties may serve as different models for essentially the same variety. To identify such tropical varieties we globalize the notion of contraction that was so far defined only for $V \subset \mathbb{T}^n$.

Let $f\colon X \to Y$ be a tropical morphism between tropical varieties of the same dimension.

Definition 3.14. The map f is called a *contraction* if for every $y \in Y$ there exist a chart $U^Y_\beta \ni y$ with $\Phi^Y_\beta(U^Y_\beta) \subset V_\beta \subset \mathbb{T}^{N_\beta}$, a contraction $\delta\colon W \to V_\beta$ and an isomorphism of polyhedral complexes

$$h\colon f^{-1}(U^Y_\beta) \approx \delta^{-1}(\Phi^Y_\beta(U_\beta))$$

such that $\Phi^Y_\beta \circ f = \delta \circ h$.

Note that a composition of contractions is again a contraction. A contraction generates an equivalence relation on the class of tropical varieties: tropical manifolds X

and Y are called equivalent if they can be connected by a sequence of contractions or operations inverse to contractions.

Example 3.15. The cubic curve given by (2) and pictured on the left-hand side of Figure 2 is equivalent to the circle S^1 equipped with the tropical structure coming from $\mathbb{R}/4.5\mathbb{Z}$ (as $\mathbb{R} = \mathbb{T}^\times$ is a tropical variety and the translation by 4.5 is a tropical automorphism there is a well-defined tropical structure on the quotient).

In the same time the real number 4.5 is an inner invariant of this cubic curve. It is a tropical counterpart of the J-invariant of elliptic curves.

It is convenient to identify equivalent tropical varieties. This allows to present an arbitrary regular morphism $f\colon X \to Y$ in Definition 3.12 by a linear tropical morphism.

Proposition 3.16. *If $f\colon X \to Y$ is a regular tropical morphism then there exists a contraction $\delta\colon \tilde{X} \to X$ such that $f \circ \delta\colon \tilde{X} \to Y$ is a linear tropical morphism.*

Example 3.17. The map $\mathbb{T} \to \mathbb{T}$ defined by $x \mapsto f(x) = \text{“}x^2 + x + 1\text{”}$ is a regular morphism which is not linear. However, the map $f \circ \delta_f\colon \Gamma_f \to \mathbb{T}$ is a linear morphism as it is given by the projection of the full graph Γ_f onto the vertical axis (cf. the left-hand side of Figure 3).

There is a well-defined notion of a k-form on a tropical variety that is preserved by the tropical equivalence.

Definition 3.18. A k-form on a tropical variety X is an exterior real-valued k-form of the integer tangent vectors at every finite point $x \in X$ such that for every chart U_α this form is induced from a (constant) linear k-form on $\mathbb{R}^{N_\alpha}$.

A k-form ω on X is called globally defined if the following condition holds for every non-finite point $x \in X$. If $x \in U_\alpha$ and $\Phi_\alpha(x) \in \mathbb{T}^{N_\alpha}$ has its jth coordinate $-\infty$ then the form ω in $\mathbb{R}^{N_\alpha}$ vanishes on the kernel of the projection onto the jth coordinate hyperplane.

E.g. the only globally defined k-form on $\mathbb{TP}^n$ is the zero form. But there might be globally defined k-forms on other compact varieties. An easy example is provided by taking X to be a *tropical torus*, the quotient of $\mathbb{R}^n$ by translation from some integer lattice $\Lambda \subset \mathbb{R}^n$ of rank n. Such X is a compact tropical variety while the globally defined k-forms on X are in 1-1 correspondence with constant linear k-forms on $\mathbb{R}^n$.

If $\delta\colon \tilde{X} \to X$ is a contraction then a globally defined k-form must vanish at all vectors in the kernel of $d\delta$, thus there is a 1-1 correspondence between forms on $\tilde{X}$ and X.

4. Tropical intersection theory

4.1. Cycles in X. The notion of an integer piecewise-linear polyhedral complex may be extended to include not only complexes in $\mathbb{T}^n$, but also complexes in an arbitrary

tropical variety X. We say that $B \subset X$ is an integer piecewise-linear polyhedral complex if for every chart $U_\alpha \subset X$ there exists an integer piecewise-linear polyhedral complex $B_\alpha \subset V_\alpha \subset \mathbb{T}^{N_\alpha}$ such that $\Phi_\alpha(B) \subset B_\alpha$.

As the overlapping maps preserve convex polyhedra we have a well-defined notion of a maximal facet on B and thus can consider *weighted* integer piecewise-linear polyhedral complexes in X.

Definition 4.1. A *k-cycle* B in a tropical variety X is a k-dimensional integer piecewise-linear polyhedral complex weighted by integer (possibly negative) numbers that satisfies Property 3.3 in every chart $\Phi_\alpha\colon U_\alpha \to V_\alpha \subset \mathbb{T}^{N_\alpha}$ of X. Accordingly, the codimension of B is $n-k$.

All k-cycles in X form a group by taking unions.

Remark 4.2. In this definition we excluded k-cycles with boundary components. Indeed, by our definition, every k-dimensional convex polyhedron in $\mathbb{T}^n$ is the closure of a convex polyhedron in $\mathbb{R}^n$. Thus we do not have components lying totally in $\mathbb{T}^n \setminus \mathbb{R}^n$. While it is useful also to consider cycles with boundary components, their intersection theory is more elaborate.

Proposition 4.3. *The image $f_*(B) \subset Y$ of a k-cycle B under a linear tropical morphism $f\colon X \to Y$ is a k-cycle.*

In particular, a tropical subvariety is a cycle. An important example is the *fundamental cycle* of the n-dimensional tropical variety X. It is the n-cycle where each facet of X is taken with its own weight.

4.2. Cycle intersections. One very useful feature of tropical varieties is the possibility to intersect cycles there.

Let B_1, B_2 be two cycles of codimension k_1 and k_2 in the same tropical variety X. The goal of this section is to define their *product-intersection* $B_1.B_2$ as a cycle of codimension k_1+k_2 in X.

We start from an easier case when the ambient variety X is $\mathbb{T}^n$. The set-theoretical intersection $B_1 \cap B_2$ is naturally stratified by convex polyhedra that are intersections of the (convex) faces of B_1 and B_2. It might happen that the dimension of some of these polyhedra is greater than $n-(k_1+k_2)$.

Definition 4.4. We define $B_1.B_2$ as the closure of the union of the strata of dimension $n-(k_1+k_2)$ in $B_1 \cap B_2$ equipped with certain weights which we define as follows.

- Suppose that an $(n-(k_1+k_2))$-dimensional stratum $S \subset B_1 \cap B_2$ is the intersection of two facets $F_1 \subset B_1$ and $F_2 \subset B_2$. Let $\Lambda_1, \Lambda_2 \subset \mathbb{Z}^n$ be the subgroups consisting of all integer vectors parallel to F_1 and F_2 respectively. Since by assumption S is of codimension k_1+k_2 the sublattice $\Lambda_1+\Lambda_2 \subset \mathbb{Z}^n$ is of finite index. We set the weight of S equal to the product of this index and the weights of F_1 and F_2.

- Suppose that the $(n-(k_1+k_2))$-stratum $S \subset B_1 \cap B_2$ is the intersection of perhaps smaller-dimensional faces $G_1 \subset B_1$ and $G_2 \subset B_2$. We choose a small vector $\vec{v} \in \mathbb{R}^n$ in the generic (non-rational with non-rational projections) direction and denote by $\tau_v \colon \mathbb{T}^n \to \mathbb{T}^n$ the translation by v (which, clearly, extends from $\mathbb{R}^n$ to $\mathbb{T}^n$). The face G_1 is adjacent to a finite number of facets $F_1^{(\alpha)}$ of B_1 while the face G_2 is adjacent to a finite number of facets $F_2^{(\beta)} \subset B_2$. As v is chosen to be generic the facets $F_1^{(\alpha)}$ and $\tau_v(F_2^{(\beta)})$ intersect transversely along a convex polyhedron parallel to S. Thus the weight of their intersection is already defined. We assign to S the weight equal to the sum of the weights of all such intersections (where the weight equals zero if $F_1^{(\alpha)}$ and $\tau_v(F_2^{(\beta)})$ are disjoint). Proposition 4.5 asserts that this total sum does not depend on the choice of v.

This definition is essentially the same as the definition of *stable intersection* from [25], note also similarities with the *Minkowski weights* from [6].

In the general case one can use in a certain way the contraction charts $\phi_\alpha : U_\alpha \to \mathbb{T}^n$ and the product-intersections $\phi_\alpha(B_1).\phi_\alpha(B_2)$ to define $B_1.B_2$ for an arbitrary tropical variety X.

Proposition 4.5. *The product-intersection $B_1.B_2$ is well defined. It gives an $(n-(k_1+k_2))$-dimensional cycle in X. The operation of taking the product-intersection is commutative and associative.*

4.3. Pull-backs, deformations, linear equivalence. Proposition 4.3 allows us to take the push-forward of a cycle under a linear tropical morphism. Using the product-intersection we can also define a pull-back.

Let $f \colon X \to Y$ be a linear tropical morphism and $B \subset Y$ is a k-cycle. The product $X \times B$ is a $(\dim X + k)$-cycle while the (set-theoretical) graph Γ_f of f is a $(\dim X)$-cycle in $X \times Y$. Their product-intersection $(X \times B).\Gamma_f$ is thus a $(k + \dim X - \dim Y)$-cycle in $X \times Y$. We define the pull-back of B by

$$f^*(B) = \pi^X_*((X \times B).\Gamma_f) \subset X,$$

where $\pi^X \colon X \times Y \to X$ is the projection onto X.

Any cycle $B \subset X \times Y$ can be considered as a family of cycles in X. Indeed, every $y \in Y$ defines a cycle

$$B_y = \pi^X_*(B.(X \times \{y\})) \subset X.$$

Definition 4.6. We call such a family *algebraic* and two cycles that appear in the same family with a connected B results of *deformation* of each other. Two cycles are called *linearly equivalent* if they appear in the same family with $Y = \mathbb{TP}^1$.

Proposition 4.7. *If $B_1, B_2 \subset X$ are two cycles and $B_1', B_2' \subset X$ are results of their deformation then $B_1'.B_2'$ is a result of deformation of $B_1.B_2$.*

The deformations are especially interesting in the case when X is compact as the following proposition shows. Note that a 0-cycle in X (as by our assumption X is covered by a finite collection of charts) is a finite union of points. We define the *degree* of a 0-cycle $B \subset X$ to be the sum of the weights of all points of B.

Proposition 4.8. *The degree of a* 0*-cycle in a compact tropical variety is a deformation invariant.*

We may define the *intersection number* of a collection of cycles of total codimension n as the degree of their product-intersection.

Example 4.9. All hypersurfaces of the same degree in $\mathbb{TP}^n$ (see Example 3.10) can be obtained from each other by deformation. Furthermore, they are linearly equivalent. We have the tropical Bezout theorem: the intersection number of n hypersurfaces of degree $d_1, \dots, d_n$ is equal to $\prod_{j=1}^n d_j$, cf. [31].

We can see the illustration to this theorem in Figures 1 and 2. In Figure 1 we have a line and a conic and they intersect in two distinct points. The weight of each of these intersection points is 1 as the primitive integer vectors parallel to the edges at the points of intersection form a basis of $\mathbb{Z}^2$. In Figure 2 we have two cubics that intersect at eight points. Out of these eight points seven have weight 1 while one, the point of intersection of a horizontal edge with an edge of slope 2, has weight 2. Thus the total intersection number of these two cubics equals 9.

4.4. Intersection with divisors. Taking the product-intersection simplifies in the case when one of the cycles is a (Cartier) divisor.

Definition 4.10. A *divisor* $D \subset X$ is a finite formal linear combination with integer coefficients of $(\dim X - 1)$-cycles given by an open covering $\{U_\alpha\}$ and regular functions $f_\alpha \colon \mathcal{U}_\alpha \to \mathbb{T}$ that defines $D \cap U_\alpha$ as its hypersurface.

Not every cycle of codimension 1 in X is a divisor.

Example 4.11. Let $X \subset \mathbb{T}^3$ be the hypersurface given by "$x+y+z+0$" (see Figure 4). Let $B \subset X$ be the line $\{(t, t, 0) \mid t \in \mathbb{T}\}$. It is a 1-cycle in the 2-dimensional tropical variety X. Yet it cannot be presented as a hypersurface near the point $(0, 0, 0)$ which is the vertex of X.

If $B \subset X$ is a k-cycle and $D \subset X$ is a divisor then in each U_α the product-intersection $D.B$ coincides with the hypersurface defined by f_α on B (i.e. with the points where "$\frac{1_\mathbb{T}}{f_\alpha}$" is not regular weighted by the order of poles of "$\frac{1_\mathbb{T}}{f_\alpha}$").

Similarly, we can define a pull-back of a divisor $D \subset Y$ under a regular morphism $h \colon X \to Y$ by taking the pull-backs of the functions f_α. As the pull-backs $(f_\alpha) \circ h$ are regular on $h^{-1}(U_\alpha)$ they define a divisor $h^*(D) \subset X$.

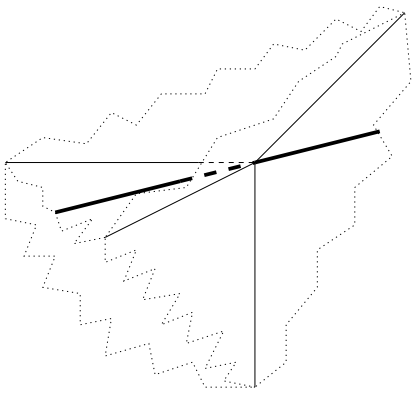


Figure 4. A cycle of codimension 1 which is not a (Cartier) divisor.

5. Tropical curves

5.1. Tropical curves as metric graphs. As in the classical case the easiest varieties to understand are curves. A tropical structure on a curve (which is topologically a graph) can be expressed by introducing a metric. Such presentation is not unlike the presentation of complex curves of negative Euler characteristic with hyperbolic surfaces. For simplicity we restrict our attention to compact tropical curves.

Recall that a *leaf* of a graph is an edge adjacent to a 1-valent vertex. We call an edge which is not a leaf a *finite edge*. Denote the set of all 1-valent vertices by Vert_1.

Definition 5.1 (cf. e.g. [1]). A *metric graph* is a finite graph Γ such that every finite edge has a prescribed positive real length. The length of all leaves is set to be $+\infty$.

This makes $\Gamma \setminus \mathrm{Vert}_1(\Gamma)$ a complete metric space (equipped with an inner metric). We denote the resulting metric by d_Γ. A homeomorphism between metric graphs is an isomorphism if it is an isometry on $\Gamma \setminus \mathrm{Vert}_1(\Gamma)$. Note that a presentation of a topological space as a graph is not unique, at our will we may introduce or erase 2-valent vertices.

Proposition 5.2. *There is a* 1-1 *correspondence between compact tropical curves and metric graphs.*

A primitive integer tangent vector at a point of a tropical curve (in every chart) has the unit length in the corresponding metric.

Proposition 5.3. *A map $f : \Gamma \to \Gamma'$ between tropical curves is a regular morphism if there exists a presentation of Γ as a graph so that for every edge $E \subset \Gamma$ there exists $n(E) \in \mathbb{N} \cup \{0\}$ such that*

$$d_{\Gamma'}(f(x), f(y)) = n(E) d_\Gamma(x, y)$$

for any $x, y \in E$.

Presentation as a metric graph is a convenient way to specify a tropical structure on a curve. The *genus* of a tropical curve Γ is its first Betti number $b_1(\Gamma)$. (This term is justified by Proposition 5.5.)

Example 5.4. Figure 5 depicts all tropical curves of genus 2. Varying the length of the edges we vary the tropical structure on the curve. For the curves in the left and in the right we have three lengths of edges, a, b and c, to vary. The middle curve is an intermediate case when one of these lengths becomes zero.

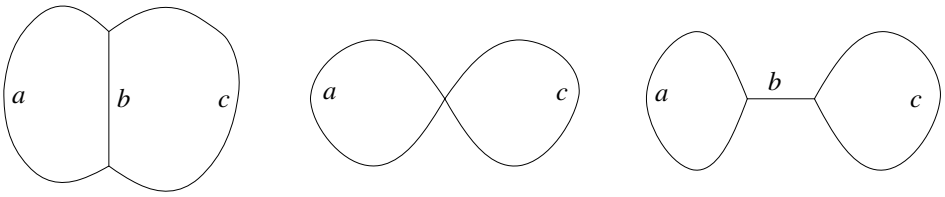


Figure 5. Tropical curves of genus 2.

This shows that the space of tropical curves of genus 2 is 3-dimensional. Furthermore, one may observe that these curves are hyperelliptic as there exists a non-trivial isometry involution for each of the three types in the picture. (The fixed points of this isometry are the midpoints of the edges for the left picture; the midpoint of the edges and the vertex for the center picture; the edge connecting the vertices and the midpoints of the other edges for the right picture.)

5.2. Jacobian varieties and the Riemann–Roch inequality

Proposition 5.5. *All* 1*-forms on a compact tropical curve* Γ *form a real vector space* $\Omega(\Gamma)$ *of dimension* $b_1(\Gamma)$*, i.e. the genus of* Γ.

Note that any 0-cycle on a curve Γ is a divisor in the sense of Definition 4.10. The divisors of degree 0 form a group Div_0 while the divisors of degree d form a homogeneous space Div_d over Div_0. We define the *Picard group* Pic_0 by taking the quotient group of Div_0 by the linear equivalence (see Definition 4.6). Similarly we define Pic_d.

A divisor is called *principal* if it is linearly equivalent to zero. Clearly, the degree of a principal divisor is 0. The classical Mittag-Leffler problem to determine whether a divisor of degree 0 is principal is answered by the Abel–Jacobi theorem. A similar answer exists also in the tropical set-up.

Note that given a 1-form ω and a path $\gamma : [0, 1] \to \Gamma$ there is a well-defined integral $\int_\gamma \omega$. Clearly, the value of this integral depends only on the relative homology class of γ. Let $\Omega^*(\Gamma)$ be the (real) dual vector space to $\Omega(\Gamma)$. Its dimension is equal to the genus g of Γ by Proposition 5.5. Each element $a \in H_1(\Gamma; \mathbb{Z})$ determines a point

of $\Omega^*(\Gamma)$. This point is given by the functional

$$\Lambda(a)\colon \omega \mapsto \int_a \omega.$$

The *Jacobian* of a tropical curve Γ is defined by

$$\mathrm{Jac}(\Gamma) = \Omega^*(\Gamma)/\Lambda(H_1(\Gamma;\mathbb{Z})).$$

The Jacobian is an example of a tropical torus. It is homeomorphic to $(S^1)^g$ and carries the structure of a tropical variety (which depends on the lattice $\Lambda(H_1(\Gamma;\mathbb{Z}))$.

To define the *Abel–Jacobi* map

$$\alpha\colon \mathrm{Pic}_0(\Gamma) \to \mathrm{Jac}(\Gamma) \tag{4}$$

we take any 1-chain C whose boundary is a divisor in the equivalence class $p \in \mathrm{Pic}_0$ and set $\alpha(p)$ to be the functional $\omega \mapsto \int_C \omega$.

Proposition 5.6 (Tropical Abel–Jacobi theorem). *The Abel–Jacobi map* (4) *is a well-defined bijection.*

This gives the structure of a tropical variety on Pic_0 as well as on Pic_d (which is a homogeneous space over Pic_0). We have the tautological map

$$\mathrm{Sym}^d(\Gamma) \to \mathrm{Pic}_d(\Gamma)$$

defined by taking the equivalence class.

As in the classical case this map is especially interesting in the case $d = g - 1$. The image of this map in this case (as well as all its translates in $\mathrm{Pic}_0 = \mathrm{Jac}$) is called the Θ-*divisor*. There is a tropical counterpart of the Riemann theorem stating that the Θ-divisor is given by a θ-function, see [23].

Figure 6 sketches the Θ-divisor in the case of genus 2. In this case it is isomorphic to the curve Γ itself. The Jacobian in Figure 6 is obtained by identifying the opposite sides of the dashed parallelogram.

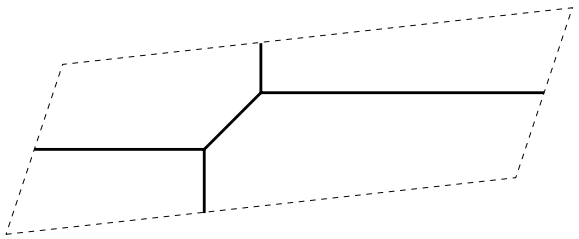


Figure 6. Θ-divisor for genus 2.

The divisor is called *effective* if the weights at all its points are positive. The Riemann–Roch theorem allows to find the dimension of all effective divisors linearly equivalent to a given one. (Alternatively, this number plus one can be interpreted as the dimension of the space of sections of the line bundle corresponding to a given divisor.) Tropically, we have the Riemann–Roch theorem in form of an inequality.

Proposition 5.7 (The tropical Riemann–Roch inequality). *The dimension of the space of effective divisors in the equivalence class $p \in \mathrm{Pic}_d$ is at least $d - g$.*

5.3. The canonical class, regular and superabundant curves in X. We have the Chern classes for compact tropical varieties which are easy to compute (as a tropical variety is already parallelized by the integer affine structure on its facets).

Let X be an n-dimensional compact tropical variety. We have the natural stratification of the faces of X by their dimension. Furthermore, there is the *boundary stratification* of X. We say that a face $F \subset X$ is in the kth boundary stratum if there is a chart $U_\alpha \ni x$ where $\Phi_\alpha(F)$ is contained in an intersection of k out of N_α tropical coordinate hyperplanes of $\mathbb{T}^{N_\alpha}$ and not contained in the intersection of any $(k+1)$ coordinate hyperplanes.

A face in the 0th boundary stratum is called a *finite* face. Clearly a face in the kth boundary stratum has dimension at most $(n-k)$.

The kth Chern class of X is an $(n-k)$-cycle that is the linear combination of all $(n-k)$-dimensional faces of X. Here the strata of boundary codimension k are taken with the weight equal to $(-1)^k$ times some positive number depending only on the local geometry near a point x in the relative interior of F (this weight can be computed inductively from a local presentation of a neighborhood of x by a contraction $\delta\colon V \to \mathbb{T}^n$).

Here we give the computation of weights only in the case of $-c_1$. This is the $(n-1)$-cycle in X that consists of all finite $(n-1)$-faces F taken with the weight equal to the number of adjacent facets to F minus 2 and all $(n-1)$-faces in the 1-boundary stratum taken with the weight -1.

Definition 5.8. The *canonical class* K is the $(n-1)$-cycle in X equal to $-c_1$.

A *parameterized tropical curve* in X is a linear tropical morphism $h\colon \Gamma \to X$. One may use Proposition 5.7 to compute the dimension in which the map h varies (we allow to deform both h and the tropical structure on Γ).

The adjunction formula exists also in the tropical geometry. As in this talk we have not defined tropical vector bundles in general and, in particular, the normal bundles, we state it only in the case when h is an embedding and X is a compact surface. We have

$$h_*([\Gamma]).h_*([\Gamma]) = 2g - 2 - K.h_*([\Gamma]). \tag{5}$$

Here $[\Gamma]$ stands for the fundamental 1-cycle in Γ.

Example 5.9. Consider the compactification of Example 4.11. Let $X \subset \mathbb{TP}^3$ be the closure of the hypersurface $x + y + z + 0$ and $h\colon \mathbb{TP}^1 \to X$ be the linear tropical morphism whose image is the closure of B. The two infinite ends of B contribute $+1$ each to its self-intersection, thus the self-intersection contribution of B at $(0, 0, 0)$ is -1 (see Figure 4)

Proposition 5.10. *The map $h\colon \Gamma \to X$ varies in a family of dimension at least*

$$\mathrm{vdim_h} = K.h_*([\Gamma]) + (n-3)(1-g)$$

if we allow to deform both h and Γ. The number vdim *is called the* virtual dimension *of the deformations of h.*

Definition 5.11. The parameterized curve h is called *regular* if the local dimension of all its deformations is equal to vdim_h. Otherwise, if it is strictly greater than vdim_h, the curve h is called *superabundant*.

E.g. the curve h will necessarily be superabundant if $h(\Gamma)$ contains a loop that is contained in a proper affine-linear subspace of a facet of X.

5.4. Tropical moduli spaces. Let Γ be a compact tropical curve of genus g. Let us choose k distinct points $x_1, \ldots, x_k \in \Gamma$. We call x_j the *marked points*. By replacing Γ with an equivalent tropical curve if needed, we may assume that Γ has exactly k 1-valent vertices which coincide with $x_1, \ldots, x_k$. Such presentation exists unless $g = 0$ and $k < 2$ and it is unique up to isomorphism. Thus we may restrict our attention only to such models.

We denote by $\mathcal{M}_{g,k}$ the space of all tropical curves of genus g with k marked points up to equivalence.

A tropical curve with marked points is determined by its combinatorial type and the lengths of its finite edges. E.g. the tropical curve from Figure 7 is determined by the lengths a and b. The length of the finite edges can be used to enhance $\mathcal{M}_{g,k}$ with

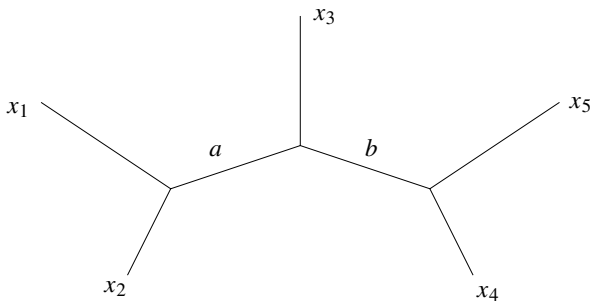


Figure 7. A rational curve with 5 marked points.

a tropical structure. The only problem is that we may only identify the edges within the same combinatorial type of the curve.

One can easily solve this problem if $g = 0$. Let us introduce a global function Z_{x_i,x_j} on $\mathcal{M}_{0,k}$ for any pair of marked points $\{x_i, x_j\}$. We set Z_{x_i,x_j} equal to the total length of finite edges between x_i and x_j. E.g. for the curve $\Gamma \in \mathcal{M}_{0,5}$ from Figure 7 we have $Z_{x_1,x_2}(\Gamma) = 0$, $Z_{x_1,x_3}(\Gamma) = a$ and $Z_{x_1,x_5}(\Gamma) = a + b$.

Proposition 5.12. *The functions Z_{x_i,x_j} define an embedding*

$$\mathcal{M}_{0,k} \subset \mathbb{R}^{\frac{k(k-1)}{2}}.$$

Its image is a tropical subvariety.

Remark 5.13. This embedding is related to the Plücker coordinates on the Grassmannian $G_{2,k}$, see [30] for a tropical version.

Thus we see that $\mathcal{M}_{0,k}$ has a natural structure of a tropical variety. We may compactify $\mathcal{M}_{0,k}$ by allowing the lengths of some (or all) finite edges of Γ to take the value equal to $+\infty$. Such a generalized curve splits into several components: the finite points of each such component are within finite distance from each other.

Proposition 5.14. *The resulting space $\overline{\mathcal{M}}_{0,k}$ is a compact tropical variety.*

This compactification is a tropical counterpart of the Deligne–Mumford compactification.

If $g > 0$ one may use similar arguments to show that $\mathcal{M}_{g,k}$ is a tropical orbifold which can be compactified to a compact tropical orbifold $\overline{\mathcal{M}}_{g,k}$.

5.5. Stable curves in X. As in the classical case we call a parameterized tropical curve $h\colon \Gamma \to X$ in a compact n-dimensional tropical variety X with compact Γ *stable* if there are no infinitesimal automorphisms of h, i.e. if the number of isomorphisms Φ of Γ such that $h = h \circ \Phi$ is finite.

All deformations of a regular stable curve h locally form a tropical variety of dimension $K.h_*([\Gamma]) + (n-3)(1-g)$. Let us fix some class β of stable curves to X that is closed with respect to regular curve deformations. The class β can be given e.g. by prescribing the intersection number with all $(n-1)$-cycles in X.

Denote by $\mathcal{M}^{\beta}_{g,k}(X)$ the space of stable curves of genus g with k marked points in the class. In many cases the space $\mathcal{M}^{\beta}_{g,k}(X)$ can be compactified to a compact tropical variety $\overline{\mathcal{M}}^{\beta}_{g,k}(X)$. This holds for instance if $g = 0$, $X = \mathbb{TP}^n$ and β is formed by curves of degree d. Another instance is if $X = \mathbb{TP}^2$, there are no restrictions on g and β is formed by topological immersions of degree d. These are the two principal cases for our enumerative applications. More generally we may assume that X is any compact toric variety, but then we have to impose the additional constraint on β that it does not contain curves with components lying totally in the boundary divisors of X.

Furthermore, in these cases we have the *evaluation* map

$$\mathrm{ev}_j\colon \overline{\mathcal{M}}^{\beta}_{g,k}(X) \to X,$$

$\mathrm{ev}_j(h) = h(x_j)$ as well as the maps $\mathrm{ft}\colon \overline{\mathcal{M}}^{\beta}_{g,k}(X) \to \overline{\mathcal{M}}_{g,k}$, $\mathrm{ft}(h) = \Gamma$, and the maps $\pi_j\colon \overline{\mathcal{M}}^{\beta}_{g,k}(X) \to \overline{\mathcal{M}}^{\beta}_{g,k-1}(X)$ "forgetting" the marked point x_j. These maps are linear tropical morphisms.

This allows to set up a tropical framework for the Gromov–Witten theory. E.g. given a collection of cycles in X we may take their pull-backs and then take their intersection number in $\overline{\mathcal{M}}^{\beta}_{g,k}(X)$.

Many reasonings in the Gromov–Witten theory can be literally repeated in this tropical set-up. A good example is the WDVV-relation, cf. [7]. As in the classical case the stable curves in $\overline{\mathcal{M}}^{\beta}_{0,k}(X) \setminus \mathcal{M}^{\beta}_{0,k}(X)$ must consist of several components.

Remark 5.15. Moduli spaces of higher-dimensional tropical varieties is a very interesting, but much more difficult subject. Already the tropical K3-surfaces form a very sophisticated geometric object, see [14] and [9].

6. Tropical curves in $\mathbb{R}^n$, their phases and amoebas

Let $V \subset (\mathbb{C}^\times)^n$ be an algebraic variety. Its *amoeba* (see [8]) is the set $\mathcal{A} = \mathrm{Log}(V) \subset \mathbb{R}^n$, where $\mathrm{Log}(z_1, \dots, z_n) = (\log|z_1|, \dots, |z_n|)$. Similarly we may consider the map

$$\mathrm{Log}_t\colon (\mathbb{C}^\times)^n \to \mathbb{R}^n$$

corresponding to taking the logarithm with the base $t > 1$.

Amoebas themselves have proved to be a very useful tool in several areas of mathematics, see e.g. [4], [16], [17], [22], [26], [27]. However, for the purposes of this talk we only use them as an intermediate link between the classical and tropical geometries.

Definition 6.1. The curve $h\colon \Gamma \to \mathbb{R}^n$ is called *classically realizable* if there exist a small regular neighborhood $U \supset B$ in $\mathbb{R}^n$, a retraction $\rho\colon U \to B$, a regular family of holomorphic maps $H_t\colon C_t \to (\mathbb{C}^\times)^n$ for a family of Riemann surfaces C_t defined for all sufficiently large positive $t \gg 1$ and smooth maps $\lambda_t\colon C_t \to \Gamma$ such that

- $h \circ \lambda_t = \rho \circ \mathrm{Log}_t \circ H_t$;
- the genus of C_t coincides with the genus of Γ;
- the number of punctures of C_t coincides with the number of ends of Γ.

The family $H_t\colon C_t \to (\mathbb{C}^\times)^n$ is called an *approximating family* of h.

Proposition 6.2. *If $h\colon \Gamma \to \mathbb{R}^n$ is a tropical curve approximated by H_t then for a sufficiently large t the inverse image $\lambda_t^{-1}(p)$ is a smooth circle for every p inside an edge of Γ while the inverse image $\lambda_t^{-1}(W)$ is diffeomorphic to a sphere with u punctures for a small connected neighborhood W of a vertex of valence u.*

In particular, if Γ is 3-valent then λ_t defines a pair-of-pants decomposition. In turn, this pair-of-pants decomposition determines a point in the boundary of the classical Deligne–Mumford space $\overline{\mathcal{M}}^{\mathbb{C}}_{g,k}$. This point is the limit of the Riemann surfaces C_t.

See [17] for a generalization of the pair-of-pants decomposition for the case of higher-dimensional hypersurfaces .

Remark 6.3. Classical realizability of tropical varieties in $\mathbb{R}^n$ is closely related to their presentation by *non-Archimedean amoebas*, the images of algebraic varieties in $(K^\times)^n$ under the coordinatewise valuations (as defined by Kapranov [10]). Here K is an algebraically closed field with a non-Archimedean valuation $\mathrm{val}\colon K^\times \to \mathbb{R}$. See [29] for an account of what is known on such presentations.

To formulate the realizability theorem in full generality we need to define the *phases* for Γ. We start from their definition in a model case.

Let $\Gamma_k \subset \mathbb{R}^k$ be the tropical curve consisting of $(k+1)$ rays emanating from $0 \in \mathbb{R}^k$ in the directions $(-1,\dots,0),\dots,(0,\dots,-1)$ and $(1,\dots,1)$. The tautological embedding $\Gamma_k \subset \mathbb{R}^k$ is easily realizable. For an approximating family we can take $H_t = L_k \subset (\mathbb{C}^\times)^n \subset \mathbb{CP}^k$, where L_k is a line with $(k+1)$ ends in $(\mathbb{C}^\times)^n$. The choice of L_k up to a multiplication by a point of $(\mathbb{R}_+)^k$ is called the phase of Γ_k.

Locally, near any point $x \in \Gamma$ the map h coincides with the map

$$\Gamma_k \subset \mathbb{R}^k \xrightarrow{A+c} \mathbb{R}^n$$

near $0 \in \Gamma_k$ for a linear map A defined over $\mathbb{Z}$ and $c \in \mathbb{R}^n$, where $(k+1)$ is the valence of x. The linear map A can be exponentiated to a multiplicative linear map $a\colon (\mathbb{C}^\times)^k \to (\mathbb{C}^\times)^n$. The phase σ_U of Γ at a small neighborhood $U \ni x$ is defined to be the equivalence class of $\xi a(L_k) \subset (\mathbb{C}^\times)^n$, $\xi \in (\mathbb{C}^\times)^n$, up to multiplication by an element of $(\mathbb{R}_+)^n$. Two local phases are called *compatible* if they agree on the intersection of the neighborhoods.

We say that H_t approximates Γ *with the local phase* σ_U if $H_t(C_t) \cap \mathrm{Log}_t^{-1}(U)$ converges (in the Hausdorff metric on compacts in $(\mathbb{C}^\times)^n$) to $\xi t^C a(L_k) \in \sigma_U$.

Theorem 1. *Any regular tropical curve $h\colon \Gamma \to \mathbb{R}^n$ equipped with any compatible system of phases is classically realizable.*

Cf. [19] and [28] for the special case $n = 2$ with no restriction on the genus, and [20] and [24] for the special case of genus 0 with no restriction on n. The proof in the general case (though in a somewhat different language) is contained in [2] (cf. also [29]).

Remark 6.4. Regularity is a necessary condition in Theorem 1, it is easy to construct an example of a non-realizable superabundant curve even for a topological immersion of an elliptic curve in $\mathbb{R}^3$, see e.g. [18].

7. Applications

One of the greatest advantages of tropical geometry is that most classical problems become much simpler after their tropicalization (if such a tropicalization exists!). This simplicity comes from the piecewise-linear nature of the tropical objects. Indeed, once we fix the combinatorial type of the data, a tropical problem becomes linear. E.g. if there are finitely many solutions to a problem then in every combinatorial type we will have a unique solution or no solutions at all.

This allows one to find (at least) an algorithmic answer to a tropical problem. Sometimes one can show that the answer to a classical problem and its tropicalization must coincide and obtain the answer to the classical problem in this way. In this last section we list some examples of such problems.

7.1. Complex geometry. Let $g \geq 0$ and $d \geq 1$ be integers. Fix a collection $\mathcal{Z} = \{z_j\}_{j=1}^{3d-1+g}$ of points in $\mathbb{CP}^2$ in general position. There are finitely many holomorphic curves of genus g and degree d passing through $\mathcal{Z}$. Let $N^{\mathbb{C}}_{g,d}$ be their number.

Such set-up can be almost literally repeated in the tropical framework. Fix a collection $\mathcal{X} = \{x_j\}_{j=1}^{3d-1+g}$ of points in $\mathbb{TP}^2$ in general position (see [19]). Again, it can be shown that there are finitely many tropical curves $h\colon \Gamma \to \mathbb{TP}^n$ of genus g and degree d passing through x_j. But in the tropical set-up these curves come with natural positive integer multiplicities $m(h)$ not necessarily equal to 1.

Tropical curves of genus g and degree d that pass through $x_j \in \mathbb{TP}^2$ form a codimension 1 cycle in $\bar{\mathcal{M}}^d_{g,d}(\mathbb{TP}^2)$. We set $m(h)$ to be the weight of their product-intersection at h (in other words this is their local intersection number in h). Let $N^{\mathbb{T}}_{g,d}$ be the number of the tropical curves of degree d and genus g passing through $\mathcal{X}$ counted with multiplicity m.

Theorem 2 ([19]).

$$N^{\mathbb{C}}_{g,d} = N^{\mathbb{T}}_{g,d}.$$

Of course, there are well-known ways to compute $N^{\mathbb{C}}_{g,d}$ (see [12], [3]) as they coincide with the Gromov–Witten invariants of $\mathbb{CP}^2$. Yet Theorem 2 gives another simple and visual way to do it, see [19] for details.

Example 7.1. Figure 8 depicts rational cubic curves in $\mathbb{TP}^2$ passing through a generic configuration $\mathcal{X}$ of 8 points. There are nine curves. Eight of them have multiplicity 1. The remaining one, namely the rightmost in the middle row, has multiplicity 4 (it has

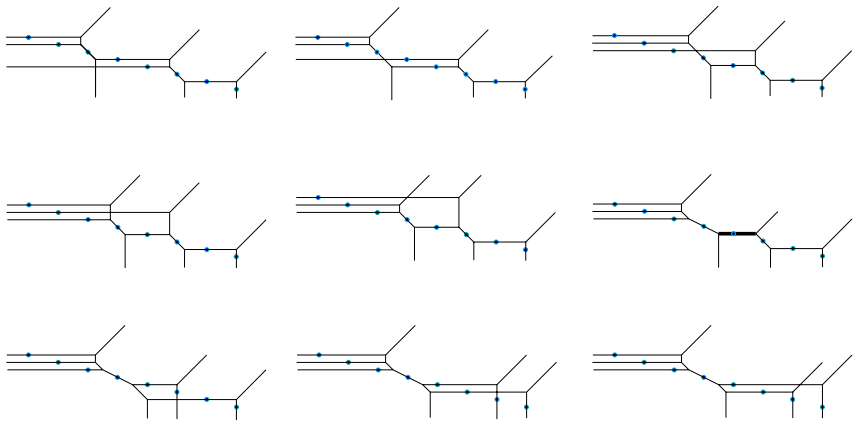


Figure 8. Rational tropical cubics via 8 generic points in the plane.

an edge of weight 2 shown by a bold line in the picture). Thus the total number of tropical curves is 12.

Should we choose a different generic configuration $\mathcal{X}$ the number of tropical curves could be different. But their total number counted with multiplicities is invariant. E.g. there exists a different choice of $\mathcal{X}$ where we have ten tropical curves, nine of multiplicity 1 and one of multiplicity 3. In fact, no other partition of 12 can appear by Example 7.3.

Remark 7.2. In [19] there is also a version of Theorem 2 for other toric surfaces. Note that if the ambient toric surface is not Fano then there is a difference between counting irreducible curves in a given homology class and computing the corresponding Gromov–Witten number (as the latter also has a contribution from curves with some boundary divisors as components). Tropical geometry has the advantage of giving a way to compute the number of irreducible curves directly without having to deal with those extra components.

The same advantage allows to apply tropical geometry for giving an algorithmic answer to another classical problem of complex geometry, namely the computation of Zeuthen's numbers, see [21]. These are the numbers of curves of degree d and genus g that pass through a collection of generic points and tangent to a collection of generic lines in $\mathbb{CP}^2$. Here the total number of points and lines in the configuration is $3d-1+g$. These Zeuthen's numbers also turn out to be equal to the corresponding tropical numbers, and the latter can be computed by a finite (though quite extensive for large genus) algorithm. As far as the author knows such computation in general is not currently accessible by non-tropical techniques.

7.2. Real geometry. The number of real curves of degree d and geometric genus g passing via a collection $\mathcal{Z} = \{z_j\}_{j=1}^{3d-1+g}$ of points in $\mathbb{RP}^2$ depends on $\mathcal{Z}$ even if we choose it to be generic. This is the feature of $\mathbb{R}$ as a non-algebraically closed field. Yet, as it was suggested in [34], one may prescribe a *sign* ± 1 to every such curve so that the sum W_d of these signs becomes invariant in the special case of $g = 0$. The number W_d is called the *Welschinger number*.

If $\mathcal{Z}$ is a generic configuration of $3d - 1$ points in $\mathbb{RP}^2$ then any real rational curve $\mathbb{R}C$ of degree d passing through $\mathcal{Z}$ is a nodal curve, i.e. all singularities of $\mathbb{R}C$ are non-degenerate double points. Over $\mathbb{R}$ there are two types of such nodes: the *hyperbolic* node, corresponding to the intersection of two real branches (given in local coordinates by $x^2 - y^2 = 0$) and the *elliptic* node, corresponding to the intersection of two complex-conjugate branches (given in local coordinates by $x^2 + y^2 = 0$). The sign of $\mathbb{R}C$ is defined as $(-1)^{e(C)}$, where $e(C)$ is the number of elliptic nodes of C.

The Welschinger number also has a tropical counterpart. To a tropical curve $h\colon \Gamma \to \mathbb{TP}^2$ of degree d and genus g passing through a configuration $\mathcal{X}$ of $3d-1+g$ points we associate its *real multiplicity* $m^{\mathbb{R}}(h)$ which is ± 1 or 0. If $h(\Gamma)$ has an edge of even multiplicity then $m^{\mathbb{R}}(h) = 0$. Otherwise we define the local real multiplicity $m_v^{\mathbb{R}}(h)$ for a vertex $v \in \Gamma$ to be $(-1)^{e_v}$, where e_v is the number of integer points in the interior of the lattice triangle such that its sides are perpendicular to the edges adjacent to v and of integer length equal to the weight of that edge. Then we define $m^{\mathbb{R}}(h) = \prod_v m_v^{\mathbb{R}}(h)$. Let $W_{g,d}^{\mathbb{T}}$ be the corresponding tropical number.

Theorem 3 ([19]).

$$W_d = W_{g,d}^{\mathbb{T}}.$$

This theorem is the only currently known way to compute the Welschinger numbers, see [11].

Example 7.3. Let us revisit Example 7.1. The real multiplicities of eight out of the nine curves in Figure 8 is 1 while the real multiplicity of the remaining one is 0. Thus we have $W_3 = 8$. For another choice of a generic configuration of eight points in $\mathbb{TP}^2$ mentioned in Example 7.1 we get nine curves with $m^{\mathbb{R}}(h) = +1$ and one with $m^{\mathbb{R}}(h) = -1$.

Note that we always have $m^{\mathbb{R}}(h) = +1$ if $m(h) = 1$; $m^{\mathbb{R}}(h) = +1$ if $m(h) = 1$ if the multiplicity is 1; and $m^{\mathbb{R}}(h) = 0$ if $m(h)$ is even. Note also that we may never get $m(h) = 2$. Since the sum of the multiplicities has to be 12 and the sum of real multiplicities has to be 8 the partitions $12 = 8 + 4 = 9 + 3$ are the only two possible partitions. In particular, there does not exist a configuration of eight points in $\mathbb{TP}^2$ such that the twelve tropical cubics will all be distinct.

Remark 7.4. There is a 3-dimensional version of the number W_d which is the number of real rational curves in $\mathbb{RP}^3$ of degree d passing through a generic configuration of $2d$ points taken with certain signs, see [35]. These numbers also have tropical counterparts and Theorem 3 extends to the 3-dimensional case providing a way to compute the real geometry numbers, see [20].

As the last application of tropical geometry we would like to mention extending the patchworking [32] to curves in real toric varieties of higher dimensions by using Theorem 1 with real phases.

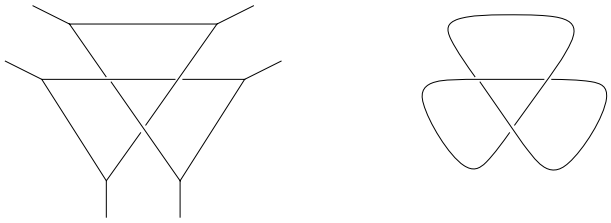


Figure 9. Real algebraic knots from tropical curves in $\mathbb{R}^3$.

Example 7.5. Let $K \subset \mathbb{R}^3$ be a knot presented as an embedded piecewise-linear circle made with k straight intervals. Then there exists an algebraic curve in $(\mathbb{R}^\times)^3$ with a unique closed component in $(\mathbb{R}_+)^3$ isotopic to K such that its complexification has genus 1 and is punctured k times. E.g. a trefoil may be presented by an elliptic curve with 6 punctures in $(\mathbb{C}^\times)^3$, see Figure 9.

To deduce this statement we perturb the broken line K to make the slopes of its intervals rational. Then we add an extra ray at every corner to get a tropical curve $\Gamma \supset K$. Finally we choose real phases for Γ so that the phases of all its bounded edges include the positive quadrant.

References

[1] Billera, L. J., Holmes, S. P., Vogtmann, K., Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27** (2001), 733–767.

[2] Bourgeois, F., A Morse-Bott approach to contact homology. Dissertation, Stanford University, 2002.

[3] Caporaso, L., Harris, J., Counting plane curves of any genus. *Invent. Math.* **131** (1998), 345–392.

[4] Einsiedler, M., Kapranov, M., Lind, D., Non-archimedean amoebas and tropical varieties. http://arxiv.org/abs/math.AG/0408311.

[5] Fukaya, K., Multivalued Morse theory, asymptotic analysis and mirror symmetry. In *Graphs and patterns in mathematics and theoretical physics,* Proc. Sympos. Pure Math. 73, Amer. Math. Soc., Providence, RI, 2005, 205–278.

[6] Fulton, W., Sturmfels, B., Intersection theory on toric varieties. *Topology* **36** (1997), 335–353.

[7] Gathmann, A., Markwig, H., Kontsevich's formula and the WDVV equations in tropical geometry. http://arxiv.org/abs/math.AG/0509628.

[8] Gelfand, I. M., Kapranov, M. M., Zelevinsky, A. V., *Discriminants, resultants, and multidimensional determinants.* Birkhäuser, Boston, MA, 1994.

[9] Gross, M., Wilson, P. M. H., Large complex structure limits of $K3$ surfaces. *J. Differential Geom.* **55** (3) (2000), 475–546.

[10] Kapranov, M., Amoebas over non-Archimedian fields. Preprint, 2000.

[11] Itenberg, I., Kharlamov, V., Shustin, E., Welschinger invariant and enumeration of real rational curves. *Internat. Math. Res. Notices* **2003** (49) (2003), 2639–2653.

[12] Kontsevich, M., Manin, Yu., Gromov-Witten classes, quantum cohomology and enumerative geometry. *Comm. Math. Phys.* **164** (1994), 525–562.

[13] Kontsevich, M., Soibelman, Ya., Homological mirror symmetry and torus fibrations. In *Symplectic geometry and mirror symmetry* (Seoul, 2000), World Sci. Publishing, River Edge, NJ, 2001, 203–263.

[14] Kontsevich, M., Soibelman, Ya., Affine structures and non-archimedean analytic spaces. In *The unity of mathematics*, Progr. Math. 244, Birkhäuser, Boston, MA, 2006, 321–385.

[15] Litvinov, G. L., The Maslov dequantization, idempotent and tropical mathematics: a very brief introduction. In *Idempotent mathematics and mathematical physics*, Contemp. Math. 377, Amer. Math. Soc., Providence, RI, 2005, 1–17.

[16] Mikhalkin, G., Real algebraic curves, the moment map and amoebas. *Ann. of Math.* (2) **151** (1) (2000), 309–326.

[17] Mikhalkin, G., Decomposition into pairs-of-pants for complex algebraic hypersurfaces. *Topology* **43** (5) (2004), 1035–1065.

[18] Mikhalkin, G., Amoebas of algebraic varieties and tropical geometry. In *Different faces of geometry,* Int. Math. Ser. (N. Y.), Kluwer/Plenum, New York 2004, 257–300.

[19] Mikhalkin, G., Enumerative tropical algebraic geometry in $\mathbb{R}^2$. *J. Amer. Math. Soc.* **18** (2) (2005), 313–377.

[20] Mikhalkin G., Rational tropical curves in $\mathbb{R}^n$. To appear.

[21] Mikhalkin, G., Zeuthen's numbers for toric surfaces via tropical geometry. To appear.

[22] Mikhalkin, G., Rullgard, H., Amoebas of maximal area. *Internat. Math. Res. Notices* **2001** (9) (2001), 441–451.

[23] Mikhalkin, G., Zharkov, I., Tropical curves, their Jacobians and Theta-functions. To appear.

[24] Nishinou, T., Siebert, B., *Toric degenerations of toric varieties and tropical curves*. http://arxiv.org/abs/math.AG/0409060.

[25] Richter-Gebert, J., Sturmfels, B., Theobald, Th., First steps in tropical geometry. In *Idempotent mathematics and mathematical physics,* Contemp. Math. 377, Amer. Math. Soc., Providence, RI, 2005, 289–317.

[26] Passare, M., Rullgard, H., Amoebas, Monge-Ampère measures, and triangulations of the Newton polytope. *Duke Math. J.* **121** (3) (2004), 481–507.

[27] Passare, M., Sadykov, T., Tsikh, A., Singularities of hypergeometric functions in several variables. *Compositio Math.* **141** (3) (2005), 787–810.

[28] Shustin, E., Patchworking singular algebraic curves, non-Archimedean amoebas and enumerative geometry. http://arxiv.org/abs/math.AG/0211278.

[29] Speyer, D., Tropical geometry. Dissertation, University of California, Berkeley, 2005.

[30] Speyer, D., Sturmfels, B., The tropical Grassmannian. *Adv. Geom.* **4** (3) (2004), 389–411.

[31] Sturmfels, B., *Solving systems of polynomial equations.* CBMS Reg. Conf. Ser. Math. 97, Amer. Math. Soc., Providence, RI, 2002.

[32] Viro, O. Ya., Gluing of algebraic hypersurfaces, smoothing of singularities and construction of curves. In *Proc. Leningrad International Topological Conference* (Leningrad, 1982), Nauka, Leningrad 1983, 149–197.

[33] Viro, O. Ya., Dequantization of real algebraic geometry on logarithmic paper. In *European Congress of Mathematics* (Barcelona, 2000), Vol. I, Progr. Math. 201, Birkhäuser, Basel 2001, 135–146.

[34] Welschinger, J.-Y., Invariants of real rational symplectic 4-manifolds and lower bounds in real enumerative geometry. *C. R. Math. Acad. Sci. Paris* **336** (4) (2003), 341–344.

[35] Welschinger, J.-Y., Spinor states of real rational curves in real algebraic convex 3-manifolds and enumerative invariants. http://arxiv.org/abs/math.AG/0311466.

Department of Mathematics, University of Toronto, Toronto ON M5S 2E4, Canada
E-mail: mikha@math.toronto.edu

Embedded minimal surfaces

William P. Minicozzi II*

Abstract. The study of embedded minimal surfaces in $\mathbb{R}^3$ is a classical problem, dating to the mid 1700s, and many people have made key contributions. We will survey a few recent advances, focusing on joint work with Tobias H. Colding of MIT and Courant Institute, and taking the opportunity to focus on results that have not been highlighted elsewhere.

Mathematics Subject Classification (2000). Primary 53A10; Secondary 53C42.

Keywords. Minimal surfaces, differential geometry, mean curvature.

1. Introduction

An immersed surface Σ in $\mathbb{R}^3$ is said to be *minimal* if it has zero mean curvature and is *embedded* if the immersion is injective. The study of embedded minimal surfaces in $\mathbb{R}^3$ is a classical problem, dating to the mid 1700s, and many people have made key contributions.

Many of the recent results have been surveyed elsewhere and we will take the opportunity to highlight results that have not been as well covered, concentrating on recent joint work with Tobias H. Colding of MIT and the Courant Institute. We will also briefly cover recent important results of W. Meeks and H. Rosenberg and of W. Meeks, J. Perez, and A. Ros. We refer to the following surveys for other perspectives:

- For more on the structure of properly embedded minimal surfaces, see the joint expository article [11] with Tobias H. Colding as well as the surveys [27] of W. Meeks and J. Perez, [35] of J. Perez, [36] of H. Rosenberg, as well as the joint surveys [14], [16], and [17] with Tobias H. Colding.

- For the construction of embedded minimal surfaces, see the surveys [22] of D. Hoffman, M. Weber, and M. Wolf, [23] of N. Kapouleas, and [38] of M. Traizet.

- For properness of minimal surfaces and the Calabi–Yau Conjectures, see the paper [15] as well as the surveys [24] of F. Martin, [16], [17], and [35].

*The author was partially supported by NSF Grant DMS 0405695.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

1.1. Embedded minimal surfaces of fixed genus. We have chosen to concentrate on the following central question:

- Can one compactify the space of embedded minimal surfaces of fixed genus?

Roughly speaking, we show in [8] that a sequence of embedded minimal surfaces with fixed genus has a subsequence that converges away from a singular set to a collection of parallel planes. The precise structure of the singular set and of the surfaces near the singular set depends on the topology of the surfaces. Consequently, we consider three separate cases:

1. When the surfaces are disks.
2. When the surfaces are (non-simply connected) planar domains; i.e., the case of genus zero.
3. When the surfaces have a fixed non-zero genus.

The case of disks was completed in [4], [5], [6] and [7] and plays a key role in the other two cases as well; the case of disks was surveyed in [14] and [16]. The other two cases, which were completed in [8], will be one of the focal points of this survey.

A key step in the compactness results for embedded minimal surfaces of fixed genus is a structure result that describes what these surfaces look like. We have chosen to focus on the compactness theorems rather than the underlying structure results, largely because it serves as a unifying theme and allows us to simplify some of the statements. Roughly speaking, two main structure theorems for (non-simply connected) embedded minimal planar domains from [8] are:

- Any such surface *without* small necks can be obtained by gluing together two oppositely-oriented double spiral staircases.
- Any such surface *with* small necks can be decomposed into "pairs of pants" by cutting the surface along a collection of short curves. After the cutting, we are left with graphical pieces that are defined over a disk with either one or two sub-disks removed (a topological disk with two sub-disks removed is called a *pair of pants*).

Both of these structures occur as different extremes in the two-parameter family of minimal surfaces known as the Riemann examples.

1.2. Embedded minimal annuli. The simplest example of a non-simply connected planar domain is of course an annulus. In [10], we obtained a precise description of what an embedded minimal annulus in a ball must look like – roughly speaking, it must look like catenoid. This illustrates a few of the ideas for the general pair of pants decomposition of [8] in a relatively simple setting. This description can be thought of as an effective version of the main theorems of [18] and [13]; i.e., [10] applies to an annulus Σ with $\partial\Sigma \subset \partial B_r(0)$ and as r goes to infinity we recover the results of [18] and [13].

1.3. Properness and removable singularities. The next result that we will highlight is the proof of "properness" in [9]. This properness was used in [7] to analyze a neighborhood of each singular point, showing that an entire neighborhood is foliated by limit planes. This can be viewed as a removable singularity theorem for minimal laminations. The proof of properness in [9] works only in the global case where we have a sequence of embedded minimal disks in a sequence of expanding balls whose radii tend to infinity – the local case is where the disks are in a fixed ball. Perhaps surprisingly, it turned out that properness can fail in the local case: In the local case, we can get limits with non-removable singularities. One local example with non-removable singularities is constructed in [12].

1.4. The global structure of complete embedded minimal surfaces in $\mathbb{R}^3$. As mentioned above, there have been many important recent developments in the field. We will survey two of these where the results of [4]–[8] play a role:

- The uniqueness of the helicoid proven by W. Meeks and H. Rosenberg in [31].
- The curvature bound for embedded minimal planar domains with bounded horizontal flux proven by W. Meeks, J. Perez, and A. Ros in [28].

The uniqueness of the helicoid solved a long-standing problem that was largely considered unapproachable until recently and also has many applications. We will sketch the proof and explain how the lamination theorem and one-sided curvature estimate played a key role.

The curvature bound of [28] was the key step in solving an old conjecture of J. Nittsche and an important step for understanding the moduli space of embedded minimal planar domains. We will explain the result, give an idea why it should be true, and explain how the compactness theorems of [8] play a role in the proof.

We should point out that there is a key distinction between these two results and the other results that we have discussed: these results both use in an essential way that the surfaces are complete and without boundary.

2. Minimal surfaces

An immersed surface $\Sigma \subset \mathbb{R}^3$ is *minimal* if it is a critical point for area, i.e., if it has zero mean curvature. The *mean curvature* is the trace of the second fundamental form A; recall that the eigenvalues of A are called the *principal curvatures*. Our surface Σ will always be embedded and will have a well-defined unit normal $\boldsymbol{n}$. The map

$$\boldsymbol{n}\colon \Sigma \to \mathbb{S}^2 \tag{1}$$

is called the *Gauss map*. Note that A is the differential of the Gauss map.

Observe that if $\Sigma \subset \mathbb{R}^3$ is minimal, then so is every rigid motion of Σ. Furthermore, so is a dilation of Σ, i.e., so is the surface

$$\lambda\,\Sigma = \{\lambda\, x \mid x \in \Sigma\}. \tag{2}$$

This is because dilating Σ by λ dilates the second fundamental form by λ^{-1}.

Note that minimal surfaces are not necessarily area-minimizing. A surface is *stable* if it satisfies the second derivative test; obviously, area-minimizing surfaces are stable.

2.1. Classical minimal surfaces. The simplest example of a minimal surface is a flat plane (where the unit normal is constant and, hence, where $A = 0$).

The only non-trivial rotationally invariant minimal surface is the *catenoid* (discovered in 1776), i.e., the minimal surface in $\mathbb{R}^3$ parametrized by

$$(\cosh s\ \cos t, \cosh s\ \sin t, s) \quad \text{where } s, t \in \mathbb{R}. \tag{3}$$

More precisely, since dilations preserve minimality, there is a one-parameter family of catenoids (modulo rigid motions) given by

$$\lambda\,(\cosh s\ \cos t, \cosh s\ \sin t, s) \quad \text{where } s, t \in \mathbb{R}. \tag{4}$$

The *helicoid* (also discovered in 1776) is the minimal surface Σ in $\mathbb{R}^3$ parametrized by

$$(s\cos t, s\sin t, t) \quad \text{where } s, t \in \mathbb{R}. \tag{5}$$

Note that the helicoid is a "double-spiral staircase", consisting of a straight line in each horizontal plane where these lines rotate at constant speed. It can also be thought of as the union of the "graphs" of the functions θ and $\theta + \pi$ together with the vertical axis. We will make this last characterization more precise later when we introduce the notion of a multi-valued graph.

The catenoid can be thought of as "two planes glued together along a small neck." Surprisingly, by a theorem of F. Lopez and A. Ros, it is impossible to glue together any other *finite* number of planes to get a complete properly embedded minimal planar domain. However, the *Riemann examples* (constructed by Riemann around 1860) give a periodic collection of horizontal planes glued together along small necks. This is actually (modulo rigid motions) a two parameter family of surfaces, where the parameters can roughly be thought of as

- the size of the necks (or injectivity radius), and
- the angle from one to the next.

As the angle goes to zero, the necks get further and further apart and the family degenerates to a collection of catenoids. As the angle goes to $\pi/2$, the necks become virtually on top of each other and the family degenerates to the union of two oppositely oriented helicoids. There are very pretty pictures of this available from David Hoffman's web page:

http://www.msri.org/about/sgp/jim/geom/minimal/library/riemann/index.html

3. Embedded minimal surfaces with fixed genus

As mentioned, we will focus on compactness theorems for a sequence $\Sigma_i \subset \mathbb{R}^3$ of embedded minimal surfaces. There are various notions of weak convergence (e.g., as currents or varifolds). However, for us, the sequence Σ_i converges to a surface Σ_∞ at a point $x \in \Sigma$ if there is a ball $B_r(x)$ so that:

- For every i sufficiently large, $B_r(x) \cap \Sigma_i$ is a (connected) graph over a (subset of) the tangent plane $T_x\Sigma_\infty$ of a function u_i.
- As $i \to \infty$, the functions u_i converge smoothly to a function u_∞ where $B_r(x) \cap \Sigma_\infty$ is the graph of u_∞.

Notice that there are two obvious necessary conditions for the sequence Σ_i to converge in this sense: The curvatures and areas of the sequence must be locally bounded.

It is not hard to see that the lack of a local area bound is not such a serious problem as long as we have embeddedness. Namely, if we have a uniform curvature bound near x, then the components of $B_r(x) \cap \Sigma_i$ are well-approximated by their tangent planes for r small. Embeddedness then implies that all of these tangent planes must be almost parallel. In particular, these components are all graphs over the same plane of functions with a uniform C^1 bound. We can use the Arzela–Ascoli theorem to pass to a subsequence that "converges" to a *collection* of minimal surfaces that do not cross. The strong maximum principle then implies that two of these limit surfaces must be identical if they touch at all, i.e., they are like the leaves of a foliation. This sort of structure is called a *lamination*.

The failure of the curvature bound is a more serious problem and will force us to allow for a singular set where the sequence simply does not converge smoothly. The simplest example of this is a sequence of rescalings $\lambda_i\,\Sigma$ with $\lambda_i \to 0$ of a fixed non-flat complete embedded minimal surface Σ. This scales the curvature by the factor λ_i^{-1} and, thus, will force the curvature to blow up at the origin. For example, a sequence of rescaled catenoids converges with multiplicity two to the punctured plane. The convergence is smooth except at 0 where $|A|^2 \to \infty$. Notice that 0 is a removable singularity for the limit.

It follows from Choi and Schoen, [1], that a similar singular compactness result holds as long as we assume a uniform bound on the total curvature:

> A subsequence converges smoothly with finite multiplicity away from a finite set of singular points; these singular points are then removable singularities for the limit surface.[1]

The situation is more complicated when there is no a priori total curvature bound. For example, if we take a sequence of rescaled helicoids, then the curvature blows up along the entire vertical axis but is bounded away from this axis. Thus we get:

[1] In fact, one can say a good deal more about the convergence and the structure of the limit; see the 1995 paper of A. Ros in the Indiana University Mathematics Journal.

- The intersection of the rescaled helicoids with a ball *away from* the vertical axis gives a collection of graphs over the plane $\{x_3 = 0\}$. As $i \to \infty$, these graphs become flat and horizontal.

- The intersection of the rescaled helicoids with a ball *centered on* the vertical axis gives a double spiral staircase, rotating faster and faster as $i \to \infty$.

In particular, the sequence of rescaled helicoids converges away from the vertical axis to a foliation by flat parallel planes.

Remark 3.1. The same thing happens when one rescales any surface asymptotic to the helicoid – such as the genus one helicoid constructed by D. Hoffman, M. Weber, and M. Wolf in [21].

If we do the same rescaling to a fixed surface in the family of Riemann examples, then we get convergence away from a line to a foliation by horizontal planes. In this case, the line is *not* perpendicular to the planes.

However, unlike the catenoid and helicoid, the Riemann examples are a two-parameter family. By choosing the two parameters appropriately, one can produce sequences of Riemann examples that illustrate both of the two structure theorems:

1. If we take a sequence of Riemann examples where the neck size is fixed and the angles go to $\frac{\pi}{2}$, then the surfaces with angle near $\frac{\pi}{2}$ can be obtained by gluing together two oppositely-oriented double spiral staircases. Each double spiral staircase looks like a helicoid. This sequence of Riemann examples converges to a foliation by parallel planes. The convergence is smooth away from the axes of the two helicoids (these two axes are the singular set where the curvature blows up).
2. Suppose now that we take a sequence of examples where the neck sizes go to zero. In this case, the surfaces can be cut along short curves into collections of graphical pairs of pants. The short curves converge to singular points where the curvature blows up and the graphical pieces converge to flat planes except at these points.

4. [8]: Compactness of embedded minimal surfaces with fixed genus

We turn next to the main compactness results of [8] for embedded minimal surfaces with fixed genus. We will restrict our discussion to the case of planar domains, i.e., when the surfaces have genus zero, to simplify things. In any case, the general case of fixed genus requires only minor changes.

In this section, $\Sigma_i \subset B_{R_i} \subset \mathbb{R}^3$ is a sequence of compact embedded minimal planar domains with $\partial\Sigma_i \subset \partial B_{R_i}$. Moreover, we will assume that $R_i \to \infty$.

The singular set $\mathcal{S}$ is defined to be the set of points where the curvature is blowing up. That is, a point y in $\mathbb{R}^3$ is in $\mathcal{S}$ for a sequence Σ_i if

$$\sup_{B_r(y)\cap\Sigma_i} |A|^2 \to \infty \quad \text{as } i \to \infty \text{ for all } r > 0. \tag{6}$$

It is not hard to see that we can pass to a subsequence so that $\mathcal{S}$ is well-defined and, furthermore, if $x \notin \mathcal{S}$, then there exists $r_x > 0$ so that

$$\sup_i \sup_{B_{r_x}(x)\cap\Sigma_i} |A| < \infty. \tag{7}$$

4.1. The finer structure of $\mathcal{S}$: Where the topology concentrates. Sequences of planar domains which are not simply connected are, after passing to a subsequence, naturally divided into two separate cases depending on whether or not the topology is concentrating at points. To distinguish between these cases, we will say that a sequence of surfaces $\Sigma_i^2 \subset \mathbb{R}^3$ is *uniformly locally simply connected* (or ULSC) if for each $x \in \mathbb{R}^3$, there exists a constant $r_0 > 0$ (depending on x) so that for every surface Σ_i

$$\text{each connected component of } B_{r_0}(x) \cap \Sigma_i \text{ is a disk.} \tag{8}$$

For instance, a sequence of rescaled catenoids where the necks shrink to zero is not ULSC, whereas a sequence of rescaled helicoids is.

Another way of locally distinguishing sequences where the topology does not concentrate from sequences where it does comes from analyzing the singular set. The singular set $\mathcal{S}$ consists of two types of points. The first type is roughly modelled on rescaled helicoids and the second on rescaled catenoids:

- A point y in $\mathbb{R}^3$ is in $\mathcal{S}_{\text{ulsc}}$ if the curvature for the sequence Σ_i blows up at y and the sequence is ULSC in a neighborhood of y.
- A point y in $\mathbb{R}^3$ is in $\mathcal{S}_{\text{neck}}$ if the sequence is not ULSC in any neighborhood of y. In this case, a sequence of closed non-contractible curves $\gamma_i \subset \Sigma_i$ converges to y.

The sets $\mathcal{S}_{\text{neck}}$ and $\mathcal{S}_{\text{ulsc}}$ are obviously disjoint and the curvature blows up at both, so $\mathcal{S}_{\text{neck}} \cup \mathcal{S}_{\text{ulsc}} \subset \mathcal{S}$. An easy argument (proposition $I.0.19$ in [6]) shows that, after passing to a further subsequence, we can assume that

$$\mathcal{S} = \mathcal{S}_{\text{neck}} \cup \mathcal{S}_{\text{ulsc}}. \tag{9}$$

Note that $\mathcal{S}_{\text{neck}} = \emptyset$ is equivalent to that the sequence is ULSC as is the case for sequences of rescaled helicoids. On the other hand, $\mathcal{S}_{\text{ulsc}} = \emptyset$ for sequences of rescaled catenoids. (These definitions of $\mathcal{S}_{\text{ulsc}}$ and $\mathcal{S}_{\text{neck}}$ are specific to the genus zero case that we are focusing on now; the definitions in the fixed genus case can be found in section 1.1 of [8].)

4.2. Compactness away from $\mathcal{S}$. If we combine the local curvature bound (7) away from $\mathcal{S}$ and a variation on the Arzela–Ascoli theorem, we can pass to a subsequence so that the Σ_i's converge away from $\mathcal{S}$ to a limit lamination $\mathcal{L}'$ of $\mathbb{R}^3 \setminus \mathcal{S}$.

The leaves of $\mathcal{L}'$ are smooth, but not necessarily complete, surfaces. To make this precise, we define the closure Γ_{Clos} of a leaf Γ of $\mathcal{L}'$ to be the union of the closures of all bounded (intrinsic) geodesic balls in Γ; that is, we fix a point $x_\Gamma \in \Gamma$ and set

$$\Gamma_{\text{Clos}} = \bigcup_r \overline{\mathcal{B}_r(x_\Gamma)}, \tag{10}$$

where $\overline{\mathcal{B}_r(x_\Gamma)}$ is the closure of $\mathcal{B}_r(x_\Gamma)$ as a subset of $\mathbb{R}^3$.

Clearly, a leaf Γ is complete if and only if $\Gamma_{\text{Clos}} = \Gamma$ and we always have that

$$\Gamma_{\text{Clos}} \setminus \Gamma \subset \mathcal{S}. \tag{11}$$

The incomplete leaves of Γ can be divided into several types, depending on how Γ_{Clos} intersects $\mathcal{S}$:

- Collapsed leaves where $\Gamma_{\text{Clos}} \cap \mathcal{S}_{\text{ulsc}}$ contains a removable singularity for Γ.
- Leaves Γ with $\Gamma_{\text{Clos}} \cap \mathcal{S}_{\text{ulsc}} \neq \emptyset$, but where Γ does not have a removable singularity. This would occur, for example, if Γ spirals infinitely into the collapsed leaf through $\Gamma_{\text{Clos}} \cap \mathcal{S}_{\text{ulsc}}$. (We show in [8] that this does not occur.)
- Leaves Γ where $\Gamma_{\text{Clos}} \setminus \Gamma \subset \mathcal{S}_{\text{neck}}$; these obviously do not occur in the ULSC case.

4.3. Disks. Before discussing the general ULSC case, it is useful to recall the case of disks. One consequence of [4]–[7] is that there are only two *local models* for ULSC sequences of embedded minimal surfaces. That is, locally in a ball in $\mathbb{R}^3$, one of following holds:

- The curvatures are bounded and the surfaces are locally *graphs* over a plane.
- The curvatures blow up and the surfaces are locally *double spiral staircases*.

Both of these cases are illustrated by taking a sequence of rescalings of the helicoid; the first case occurs away from the axis, while the second case occurs on the axis.

Using in part this local description, we were able to prove that any sequence of embedded minimal disks with curvatures blowing up has a subsequence that converges to a foliation by parallel planes. This convergence is away from a Lipschitz curve $\mathcal{S}$ that is transverse to the planes. (See the appendix for the precise statements.)

4.4. Planar domains: the general structure theorems. We will show that every sequence Σ_i has a subsequence that is either ULSC or for which $\mathcal{S}_{\text{ulsc}}$ is empty. This is the next "no mixing" theorem. We will see later that these two different cases give two very different structures.

Theorem 4.1 (No mixing theorem, [8]). *If the Σ_i's are genus zero, then there is a subsequence with either $\mathcal{S}_{\text{ulsc}} = \emptyset$ or $\mathcal{S}_{\text{neck}} = \emptyset$.*

Common for both the ULSC case and the case where $\mathcal{S}_{\text{ulsc}}$ is empty is that the limits are always laminations by flat parallel planes and the singular sets are always closed subsets contained in the union of the planes. This is the content of the next theorem:

Theorem 4.2 (Planar lamination theorem, [8]). *If the Σ_i's are genus zero and*

$$\sup_{B_1 \cap \Sigma_i} |A|^2 \to \infty, \tag{12}$$

then there exists a subsequence Σ_j, a lamination $\mathcal{L} = \{x_3 = t\}_{\{t \in \mathcal{I}\}}$ of $\mathbb{R}^3$ by parallel planes (where $\mathcal{I} \subset \mathbb{R}$ is a closed set), and a closed nonempty set $\mathcal{S}$ in the union of the leaves of $\mathcal{L}$ such that after a rotation of $\mathbb{R}^3$:

(A) *For each $1 > \alpha > 0$, $\Sigma_j \setminus \mathcal{S}$ converges in the C^α-topology to the lamination $\mathcal{L} \setminus \mathcal{S}$.*

(B) *$\sup_{B_r(x) \cap \Sigma_j} |A|^2 \to \infty$ as $j \to \infty$ for all $r > 0$ and $x \in \mathcal{S}$. (The curvatures blow up along $\mathcal{S}$.)*

4.5. Planar domains: the fine structure theorems. We will assume here that the Σ_i's are not disks (recall that the case of disks was dealt with in [4]–[7]). In particular, we will assume that for each i, there exists some $y_i \in \mathbb{R}^3$ and $s_i > 0$ so that

$$\text{some component of } B_{s_i}(y_i) \cap \Sigma_i \text{ is not a disk.} \tag{13}$$

Moreover, if the non-simply connected balls $B_{s_i}(y_i)$ "run off to infinity" (i.e., if each connected component of $B_{R_i'}(0) \cap \Sigma_i$ is a disk for some $R_i' \to \infty$), then the results of [4]–[7] apply. Therefore, after passing to a subsequence, we can assume that the surfaces are uniformly not disks, namely, that there exists some $R > 0$ so that (13) holds with $s_i = R$ and $y_i = 0$ for all i.

In view of Theorem 4.1 and the earlier results for disks, it is natural to first analyze sequences that are ULSC, so where $\mathcal{S}_{\text{neck}} = \emptyset$, and second analyze sequences where $\mathcal{S}_{\text{ulsc}}$ is empty. We will do this next.

4.6. ULSC sequences. Loosely speaking, our next result shows that when the sequence is ULSC (but not simply connected), a subsequence converges to a foliation by parallel planes away from two lines $\mathcal{S}_1$ and $\mathcal{S}_2$. The lines $\mathcal{S}_1$ and $\mathcal{S}_2$ are disjoint and orthogonal to the leaves of the foliation and the two lines are precisely the points where the curvature is blowing up. This is similar to the case of disks, except that we get two singular curves for non-disks as opposed to just one singular curve for disks.

Theorem 4.3 (ULSC compactness, [8]). *Let a sequence Σ_i, limit lamination $\mathcal{L}$, and singular set $\mathcal{S}$ be as in Theorem* 4.2. *Suppose that each Σ_i satisfies* (13) *with $s_i = R > 1$ and $y_i = 0$. If every Σ_i is ULSC and*

$$\sup_{B_1 \cap \Sigma_i} |A|^2 \to \infty, \tag{14}$$

then the limit lamination $\mathcal{L}$ is the foliation $\mathcal{F} = \{x_3 = t\}_t$ and the singular set $\mathcal{S}$ is the union of two disjoint lines $\mathcal{S}_1$ and $\mathcal{S}_2$ such that:

(C_{ulsc}) *Away from $\mathcal{S}_1 \cup \mathcal{S}_2$, each Σ_j consists of exactly two multi-valued graphs spiraling together. Near $\mathcal{S}_1$ and $\mathcal{S}_2$, the pair of multi-valued graphs form double spiral staircases with opposite orientations at $\mathcal{S}_1$ and $\mathcal{S}_2$. Thus, circling only $\mathcal{S}_1$ or only $\mathcal{S}_2$ results in going either up or down, while a path circling both $\mathcal{S}_1$ and $\mathcal{S}_2$ closes up.*

(D_{ulsc}) *$\mathcal{S}_1$ and $\mathcal{S}_2$ are orthogonal to the leaves of the foliation.*

Remark 4.4. See Appendix A for the definition of a multi-valued graph. Roughly speaking a multi-valued graph is locally a graph over a subset of a plane, but fails to be a global graph since the projection to the plane is not one-to-one.

4.7. Sequences that are not ULSC. When the sequence is no longer ULSC, one can get other types of curvature blow-up by considering the family of embedded minimal planar domains known as the Riemann examples. Recall that, modulo translations and rotations, this is a two-parameter family of periodic minimal surfaces, where the parameters can be thought of as the size of the necks and the angle from one fundamental domain to the next.

With these examples in mind, we are now ready to state our second main structure theorem describing the case where $\mathcal{S}_{\text{ulsc}}$ is empty.

Theorem 4.5 ([8]). *Let a sequence Σ_i, limit lamination $\mathcal{L}$, and singular set $\mathcal{S}$ be as in Theorem* 4.2. *If $\mathcal{S}_{\text{ulsc}} = \emptyset$ and*

$$\sup_{B_1 \cap \Sigma_i} |A|^2 \to \infty, \tag{15}$$

then $\mathcal{S} = \mathcal{S}_{\text{neck}}$ by (9) *and*

(C_{neck}) *Each point y in $\mathcal{S}$ comes with a sequence of graphs in Σ_j that converge to the plane $\{x_3 = x_3(y)\}$. The convergence is in the C^∞ topology away from the point y and possibly also one other point in $\{x_3 = x_3(y)\} \cap \mathcal{S}$. If the convergence is away from one point, then these graphs are defined over annuli; if the convergence is away from two points, then the graphs are defined over disks with two subdisks removed.*

4.8. An overview of the proofs: The ULSC case. A key point will be that the results of [4]–[7] for disks will give a sequence of multi-valued graphs in the Σ_j's near each point $x \in \mathcal{S}_{\text{ulsc}}$. Moreover, these multi-valued graphs close up in the limit to give a leaf of $\mathcal{L}'$ which extends smoothly across x. Such a leaf is said to be *collapsed*; in a neighborhood of x, the leaf can be thought of as a limit of double-valued graphs where the upper sheet collapses onto the lower. We show that every collapsed leaf is stable, has at most two points of $\mathcal{S}_{\text{ulsc}}$ in its closure, and these points are removable singularities. These results on collapsed leaves are applied first in the USLC case and then again to get the structure of the ULSC regions of the limit in general, i.e., (C2) and (D) in Theorem C.1.

Roughly speaking, there are two main steps to the proof of Theorem 4.3:

1. Show that each collapsed leaf is in fact a plane punctured at two points of $\mathcal{S}$ and, moreover, the sequence has the structure of a double spiral staircase near both of these points, with opposite orientations at the two points.
2. Show that leaves which are nearby a collapsed leaf of $\mathcal{L}'$ are also planes punctured at two points of $\mathcal{S}$. (We call this "properness".)

4.9. An overview of the proofs: The general structure. Theorem 4.5, as well as Theorem 4.1, are proven by first analyzing sequences of minimal surfaces without any assumptions on the sets $\mathcal{S}_{\text{ulsc}}$ and $\mathcal{S}_{\text{neck}}$. The precise statement of this general theorem is given in Appendix C. We will give an overview of the theorem next.

In this general case, we show that a subsequence converges to a lamination $\mathcal{L}'$ divided into regions where Theorem 4.3 holds and regions where Theorem 4.5 holds. This convergence is in $C^{1,1}$ topology away from the singular set $\mathcal{S}$ where the curvature blows up. Moreover, each point of $\mathcal{S}$ comes with a plane and these planes are essentially contained in $\mathcal{L}'$. The set of heights of the planes is a closed subset $\mathcal{I} \subset \mathbb{R}$ but may not be all of $\mathbb{R}$ as it was in Theorem 4.3 and may not even be connected. The behavior of the sequence is different at the two types of singular points in $\mathcal{S}$ – the set $\mathcal{S}_{\text{neck}}$ of "catenoid points" and the set $\mathcal{S}_{\text{ulsc}}$ of ULSC singular points. We will see that $\mathcal{S}_{\text{ulsc}}$ consists of a union of Lipschitz curves transverse to the lamination $\mathcal{L}$. This structure of $\mathcal{S}_{\text{ulsc}}$ implies that the set of heights in $\mathcal{I}$ which intersect $\mathcal{S}_{\text{ulsc}}$ is a union of intervals; thus this part of the lamination is foliated. In contrast, we will not get any structure of the set of "catenoid points" $\mathcal{S}_{\text{neck}}$. Given a point y in $\mathcal{S}_{\text{neck}}$, we will get a sequence of graphs in Σ_j converging to a plane through y. This convergence will be in the smooth topology away from either one or two singular points, one of which is y. Moreover, this limit plane through y will be a leaf of the lamination $\mathcal{L}$.

The key steps for proving the general structure theorem are the following:

1. Finding a *stable* plane through each point of $\mathcal{S}_{\text{neck}}$. This plane will be a limit of a sequence of stable graphical annuli that lie in the complement of the surfaces.
2. Finding graphs in Σ_j that converge to a plane through each point of $\mathcal{S}_{\text{neck}}$. To do this, we look in regions between consecutive necks and show that in any

such region the surfaces are ULSC. The one-sided curvature estimate will then allow us to show that these regions are graphical.

3. Using (1) and (2) we then analyze the ULSC regions of a limit. That is, we show that if the closure of a leaf in $\mathcal{L}'$ intersects $\mathcal{S}_{\text{ulsc}}$, then it has a neighborhood that is ULSC. This will allow us to use the argument for the proof of Theorem 4.3 to get the same structure for such a neighborhood as we did in case where the entire surfaces where ULSC.

The main point left in Theorem 4.5, which is not included in this general compactness theorem, is to prove that *every* leaf of the lamination $\mathcal{L}$ in Theorem 4.5 is a plane. In contrast, the general compactness theorem gives a plane through each point of $\mathcal{S}_{\text{neck}}$, but does not claim that the leaves of $\mathcal{L}'$ are planar.

Finally, since the no mixing theorem implies that Theorem 4.3 and Theorem 4.5 cover all cases, Theorem 4.2 will be a corollary of these two theorems.

5. The structure of embedded minimal annuli

We turn next to a local structure theorem for embedded minimal annuli that, roughly speaking, shows that they must look like catenoids. Namely, the main theorem of [10] proves that any embedded minimal annulus in a ball (with boundary in the boundary of the ball and) with a small neck can be decomposed by a simple closed geodesic into two graphical sub-annuli. Moreover, there is a sharp bound for the length of this closed geodesic in terms of the separation (or height) between the graphical sub-annuli. This serves to illustrate the "pair of pants" decomposition from [8] in the special case where the embedded minimal planar domain is an annulus.

The precise statement of this decomposition for annuli is:

Theorem 5.1 (Main Theorem, [10]). *There exist $\varepsilon > 0$, $C_1,\ C_2,\ C_3 > 1$ so: If $\Sigma \subset B_R \subset \mathbb{R}^3$ is an embedded minimal annulus with $\partial\Sigma \subset \partial B_R$ and $\pi_1(B_{\varepsilon R} \cap \Sigma) \neq 0$, then there is a simple closed geodesic $\gamma \subset \Sigma$ of length ℓ so that:*

- *The curve γ splits the connected component of $B_{R/C_1} \cap \Sigma$ containing it into annuli Σ^+ and Σ^-, each with $\int |A|^2 \leq 5\,\pi$.*
- *Each of $\Sigma^{\pm} \setminus \mathcal{T}_{C_2\,\ell}(\gamma)$ is a graph with gradient ≤ 1.*
- *$\ell \log(R/\ell) \leq C_3\, h$ where the separation h is given by*

$$h = \min_{x_{\pm} \in \partial B_{R/C_1} \cap \Sigma^{\pm}} |x_+ - x_-|. \tag{16}$$

Here $\mathcal{T}_s(S) \subset \Sigma$ denotes the intrinsic s-tubular neighborhood of a subset $S \subset \Sigma$.

5.1. A sketch of the proof. We will next give a brief sketch of the proof of the decomposition theorem, Theorem 5.1. The starting point is to use the hypothesis $\pi_1(B_{\varepsilon R} \cap \Sigma) \neq 0$ and a barrier argument to find a stable graph Γ_0 that is defined over an annulus and disjoint from Σ. The stable graph Γ_0 will allow us to divide Σ into two pieces, one on each side of Γ_0. To do this, we first fix a simple closed $\tilde{\gamma} \subset B_{\varepsilon R} \cap \Sigma$ that separates the two boundary components of Σ. The curve $\tilde{\gamma}$ is contained in a small extrinsic ball, but there is no a priori reason why it must be short.[2] A barrier argument using a result of Meeks and Yau then gives a stable embedded minimal annulus Γ that separates the two boundary components of Σ and where $\tilde{\gamma}$ is one component of the boundary $\partial\Gamma$ and the other component is in ∂B_R. Finally, Theorem 0.3 of [6] then implies that Γ contains the desired graph Γ_0; this should be compared with the well-known result of D. Fischer-Colbrie in the complete case.

We will see next that each half of Σ, i.e., the part above Γ_0 and the part below Γ_0, is itself a graph away from the boundary of Γ_0. This part of the argument applies more generally to an "annular end" of a minimal surface. We will prove that each half of Σ contains a graph by showing that it must contain large locally graphical pieces and then using embeddedness to see that these pieces must be global graphs (i.e., the projection down is one-to-one). This follows by combining three facts:

(1) The one-sided curvature estimate of [4]–[7] gives a scale-invariant curvature estimate for Σ's in a narrow cone about the graph Γ_0. This requires that we know that each component of Σ in balls away from the origin is a disk; this can be seen from the maximum principle.

(2) Using (1), the gradient estimate gives a narrower cone about Γ_0 where Σ is locally graphical. This is because (1) implies that the surface is well-approximated by its tangent plane and, since it cannot cross Γ_0, it must be almost parallel to Γ_0.

(3) As long as ε is small enough, each half of Σ must intersect any narrow cone about Γ_0. This was actually proven in lemma 3.3 of [9] that gave the existence of low points in a *connected* minimal surface contained on one side of a plane and with interior boundary close to this plane.[3]

Step (3) allows us to find very flat regions in Σ near Γ_0, we can then repeatedly apply the Harnack inequality to build this out into large locally graphical regions that stay inside the narrow cone about Γ_0. These locally graphical regions piece together to give a graph over an annulus; the other possibility would be to form a multi-valued graph, but this is impossible since such a multi-valued graph would be forced to spiral infinitely (since it cannot cross itself and also cannot cross the stable graph Γ_0).

[2]However, the chord arc bounds in the later paper [15] could now be used to bound the length.

[3]The argument for this was by contradiction. Namely, if there were no low points, then we would get a contradiction from the strong maximum principle by first sliding a catenoid up under the surface and then sliding the catenoid horizontally away, eventually separating two boundary components of the surface. Here the strong maximum principle is used to keep the sliding catenoids and the surface disjoint. See, for instance, corollary 1.18 in [2] for a precise statement of the strong maximum principle.

Finally, the last step of the proof is to use a blow up argument to get the precise bounds on the length of the curve γ.

5.2. Complete properly embedded minimal annuli. The decomposition of properly embedded minimal annuli given by Theorem 5.1 can be viewed as a local version of well-known global results of P. Collin, [18], and Colding and the author, [13], on annular ends.

To explain these global results, recall that Σ is said to have *finite topology* if it is homeomorphic to a closed Riemann surface with a finite number of punctures; the genus of Σ is then the genus of this Riemann surface and the number of punctures is the number of ends. It follows that a neighborhood of each puncture corresponds to a properly embedded annular end of Σ. Perhaps surprisingly at first, the more restrictive case is when Σ has more than one end. The reason for this is that a barrier argument gives a stable minimal surface between any pair of ends. Such a stable surface is then asymptotic to a plane (or catenoid), essentially forcing each end to live in a half-space. Using this restriction, P. Collin proved:

Theorem 5.2 (Main theorem, [18]). *Each end of a complete properly embedded minimal surface with finite topology and at least two ends is asymptotic to a plane or catenoid.*

In particular, such a Σ has finite total curvature and, outside some compact set, Σ is given by a finite collection of disjoint graphs over a common plane.

As mentioned above, Collin proved Theorem 5.2 by showing that an embedded annular end that lives in a half-space must have finite total curvature. [13] used the one-sided curvature estimate to strengthen this from a half-space to a strictly larger cone, and in the process give a very different proof of Collin's theorem.

Theorem 5.3 (Main theorem, [13]). *There exists $\varepsilon > 0$ so that any complete properly embedded minimal annular end contained in the cone*

$$\{x_3 \geq -\varepsilon\,(x_1^2 + x_2^2 + x_3^2)^{1/2}\} \tag{17}$$

is asymptotic to a plane or catenoid.

6. Properness and removable singularities for minimal laminations

The compactness theorems of [4]–[8] assume that the surface Σ_i has boundary $\partial\Sigma_i$ in the boundary ∂B_{R_i} of an expanding sequence of balls where R_i goes to infinity. We call this the *global* case, in contrast to the *local* case where the boundaries are in the boundary of a fixed ball ∂B_R.[4]

This distinction between the local and global cases explains why the global compactness theorem for sequences of disks does not imply the compactness theorem for

[4]One can also consider the more restrictive *complete* case where Σ_i is complete without boundary.

ULSC sequences. Namely, even though the ULSC sequence consists *locally* of disks, the compactness result for disks was in the *global* case where the radii go to infinity and hence does not apply.

In order to focus the discussion, we will explain the differences between the global and local cases for disks. The assumption that $R_i \to \infty$ is used in the compactness theorem for disks in two ways:

(1) We show that the limit lamination contains a stable leaf through each singular point. Since $R_i \to \infty$, this stable leaf is complete and, hence, a plane by the Bernstein theorem of D. Fischer-Colbrie and R. Schoen and M. Do Carmo and C. Peng.

(2) We show next that the leaves nearby a singular point must also be planes. It follows that the singular set cannot stop and all of $\mathbb{R}^3$ is foliated by planes in the limit. We call this *properness*.

The use of $R_i \to \infty$ in (1) is not really essential. The leaf would no longer have to be flat in the local case, but it would satisfy uniform estimates by R. Schoen's curvature estimate for stable surfaces, [37] (cf. [3]).

In contrast, it turns out that the use of $R_i \to \infty$ in (2) is essential. Namely, in [14], we constructed a sequence of embedded minimal disks Σ_i in the unit ball B_1 with $\partial\Sigma_i \subset \partial B_1$ where the curvatures blow up only at 0 and

$$\Sigma_i \setminus \{x_3 = 0\} \tag{18}$$

converges to two embedded minimal disks

$$\Sigma^- \subset \{x_3 < 0\}, \tag{19}$$

$$\Sigma^+ \subset \{x_3 > 0\}, \tag{20}$$

each of which spirals into $\{x_3 = 0\}$ and thus is not proper. Thus, in the example from [14], 0 is the first, last, and only point in $\mathcal{S}_{\text{ulsc}}$ and the limit lamination consists of three leaves: Σ^+, Σ^-, and the punctured unit disk $B_1 \cap \{x_3 = 0\} \setminus \{0\}$. This lamination of $B_1 \setminus \{0\}$ cannot be extended smoothly to a lamination of B_1; that is to say, 0 is not a removable singularity. This should be contrasted with the global case where every singular point is a removable singularity for the limit foliation by parallel planes. B. Dean has constructed similar examples where the singular set is an arbitrary finite set of points in the vertical axis; see [20].

6.1. A sketch of the proof of properness for disks. To explain the proof of properness in the global case for disks, we first need to see what could go wrong. Suppose therefore that the origin 0 is a singular point and $\{x_3 = 0\}$ is the corresponding limit plane. It follows from the one-sided curvature estimate that the intersection of each Σ_j with a low cone about $\{x_3 = 0\}$ consists of two multi-valued graphs for j large (the fact that there are exactly two is established in proposition II.1.3 in [7]). There are now two possibilities:

(P) The multi-valued graphs in the complement of the cone close up in the limit to a foliation.

(N-P) These multi-valued graphs converge to a collection of graphs and *at least one* multi-valued graph that spirals infinitely on one side of $\{x_3 = 0\}$.

As we saw above, the second case (N-P) can occur in the local case. We will explain why it cannot happen in the global case.

Suppose therefore that (N-P) holds and the limit contains a multi-valued graph that spirals infinitely down to the plane $\{x_3 = 0\}$; it is the graph of a multi-valued function $u(\rho, \theta)$ defined for all $\rho \geq \mathrm{e}$ and *all* $\theta > 0$. The separation $w(\rho, \theta)$ between consecutive sheets is by definition

$$w(\rho, \theta) = u(\rho, \theta + 2\pi) - u(\rho, \theta). \tag{21}$$

Since the limit is embedded and spirals downward, we must have $w < 0$. We will actually work with the conformally changed functions $\tilde{u}(x + iy) = u(\mathrm{e}^x, y)$ and $\tilde{w}(x + iy) = w(\mathrm{e}^x, y)$ that are defined on the quadrant $\{x > 1,\ y > 0\}$. The key point in the proof of properness is to show that:

(Key) The vertical flux across $\{x = 1\}$ is negative infinity.

Why (Key) *leads to a contradiction*: To see why (Key) leads to a contradiction, we need to recall more about the limit in case (N-P). Namely, we showed in [7] that there must be two multi-valued graphs spiralling together just as occurs for the helicoid. The same argument applies to both multi-valued graphs, so both have unbounded negative flux across $\{x = 1\}$, i.e., over the circle of radius e in the plane. Moreover, we also showed in [7] that these two halves can be joined together by curves in the embedded minimal disk with a uniform bound on the length of the curves. For example, the helicoid contains two infinite valued graphs and these can be connected by horizontal lines. In any case, this leads to a flux contradiction: Stokes' theorem implies that the sum of the fluxes across compact subcurves over the circle of radius e must be bounded by the length of the connecting curves. However, the length of the connecting curves is uniformly bounded and the fluxes across the other curves both go to negative infinity.

The idea of the proof of (Key): In order to keep things simple, we will pretend that u is a harmonic function; this serves to illustrate the main ideas. Since the separation w is locally the difference of two harmonic functions, w is also harmonic; hence, the conformally changed functions $\tilde{u}$ and $\tilde{w}$ are harmonic on the quadrant $\{x > 1,\ y > 0\}$. Note that $\tilde{u}$ is positive and $\tilde{w}$ is negative.

The property (Key) is now roughly equivalent to showing that

$$\int_0^\infty \frac{\partial \tilde{u}}{\partial x}(1, y)\, dy = +\infty. \tag{22}$$

It may be helpful to consider an example; the function $\tilde{u} = \pi/2 - \arctan(y/x)$ is positive, harmonic, and its multi-valued graph is an embedded infinite spiral that

accumulated to the plane $\{x_3 = 0\}$. Furthermore, is is easy to verify (22) in this case:

$$\int_0^\infty \frac{\partial \tilde{u}}{\partial x}(1, y)\, dy = \int_0^\infty \frac{y}{1+y^2}\, dy = +\infty. \tag{23}$$

To prove (22) for a general function $\tilde{u}$, first observe that Stokes' theorem gives

$$\int_1^R \frac{\partial \tilde{u}}{\partial x}(1, y)\, dy + \int_1^R \frac{\partial \tilde{u}}{\partial y}(x, 1)\, dx = \int_{\{x^2+y^2=R^2+1,\, x>1,\, y>1\}} \frac{\partial u}{\partial r}. \tag{24}$$

In [10], we used the lower bound $\tilde{u} \geq 0$ to prove that

$$\int_{\{x^2+y^2=R^2+1,\, x>1,\, y>1\}} \frac{\partial u}{\partial r} \tag{25}$$

is essentially non-negative. The reason for this is that (25) measures the logarithmic rate of growth of the average of $\tilde{u}$ on the semi-circle; if this was negative, the function $\tilde{u}$ would eventually have to become negative.

We then proved the claim (24) in [10] by using a sharp estimate for the decay of $\tilde{w}$ to show that

$$\int_1^\infty \frac{\partial \tilde{u}}{\partial y}(x, 1)\, dx = -\infty. \tag{26}$$

To explain this, observe that $\tilde{w}(x, 1)$ is nothing more than $\tilde{u}(x, 1+2\pi) - \tilde{u}(x, 1)$ and, hence, can be written as

$$\tilde{w}(x, 1) = \int_1^{1+2\pi} \frac{\partial \tilde{u}}{\partial y}(x, y)\, dy. \tag{27}$$

In particular, $\tilde{w}(x, 1) \approx 2\pi\, \frac{\partial \tilde{u}}{\partial y}(x, 1)$. We proved in [10] that the fastest possible decay for $|\tilde{w}(x, 1)|$ is c_1/x and, consequently, we get that

$$\int_1^\infty \tilde{w}(x, 1)\, dx = -\infty. \tag{28}$$

This completes the sketch of the proof. The actual argument in [10] is somewhat more complicated, but similar in flavor.

7. The uniqueness of the helicoid

The helicoid and plane are the only classical examples of properly embedded complete minimal disks in $\mathbb{R}^3$. It turns out that there is a good reason for the scarcity of examples. Namely, using the compactness theorem and one-sided curvature estimate of [4]–[7], W. Meeks and H. Rosenberg proved the uniqueness of the helicoid:

Theorem 7.1 (Main theorem, [31]). *The plane and helicoid are the only complete properly embedded simply-connected minimal surfaces in* $\mathbb{R}^3$.

This uniqueness has many applications, including additional regularity of the singular set $\mathcal{S}$. To set this us, recall that if we take a sequence of rescalings of the helicoid, then the singular set $\mathcal{S}$ for the convergence is the vertical axis perpendicular to the leaves of the foliation. In [25], W. Meeks used this fact together with the uniqueness of the helicoid to prove that the singular set $\mathcal{S}$ in Theorem B.1 is always a straight line perpendicular to the foliation.

There is an analog of Theorem 7.1 in the higher genus case. Namely, any properly embedded minimal surface with finite (non-zero) genus and one end must be asymptotic to a helicoid. Until recently, it was not known whether any such surface exists; however, the construction of the genus one helicoid in [21] suggests that there may be a substantial theory of these.

Remark 7.2. It follows from [15] that any complete embedded minimal surface with finite topology in $\mathbb{R}^3$ is automatically properly embedded. In particular, the hypothesis of properness can be removed from Theorem 7.1.

7.1. A sketch of the proof. We will give a brief overview of the proof by Meeks and Rosenberg for the uniqueness of the helicoid; we refer to the original paper [31] for the details.

The first main step in the proof of Theorem 7.1 is to analyze the asymptotic structure of a non-flat embedded minimal disk Σ, showing that it looks roughly like a helicoid. This is done in [31] by analyzing sequences of rescalings of Σ. This rescaling argument yields a sequence of embedded minimal disks which does not converge in the classical sense (there are no local area bounds). However, the lamination theorem of [4]–[7] gives that a subsequence converges to a foliation by parallel planes away from a Lipschitz curve transverse to these planes. Moreover, the lamination theorem also gives that the intersection of Σ with a cone consists of two asymptotically flat multi-valued graphs. In particular, this foliation is unique, i.e., does not depend on the choice of subsequence. After possibly rotating $\mathbb{R}^3$, we can assume that the limit foliation is by horizontal planes (i.e., level sets of x_3).

The second main step is to show that the height function x_3 together with its harmonic conjugate (which we will denote by x_3^*) give global isothermal coordinates on Σ. This step is crucial since it reduces the problem to analyzing the potential Weierstrass data on the plane. There are two key components to getting these global coordinates, both of independent interest. First, one must show that ∇x_3 does not vanish on Σ – i.e., that the Gauss map misses the north and south poles. Second, one must show that the map (x_3, x_3^*) is proper – i.e., that x_3^* goes to infinity as we go out horizontally. Both of these steps strongly use the asymptotic structure established in the first step.

The third main step is to analyze the Weierstrass data in the conformal coordinates (x_3, x_3^*). In these coordinates, the only unknown is a meromorphic function g that is the stereographic projection of the Gauss map of Σ. Since the Gauss map was already shown to miss the north and south poles, the function g can be written as

$g = e^f$ for an entire holomorphic function f. Meeks and Rosenberg then use a Picard type argument to show that f must be linear. The key in this argument is to analyze the inverse images of horizontal circles in $\mathbb{S}^2$ under the Gauss map $\boldsymbol{n}$, using rescaling arguments and the compactness theory of [4]–[7] to control the number of components. Finally, every linear function f gives rise to a surface in the associate surface family of the helicoid, but the actual helicoid is the only one of these that is embedded.

8. Quasiperiodicity of properly embedded minimal planar domains

We turn next to recent results of W. Meeks, J. Perez, and A. Ros on the structure of *complete* properly embedded minimal planar domains with infinitely many ends in $\mathbb{R}^3$. They have obtained many important results on these surfaces in the series of papers [28]–[30]; we have chosen to focus on the main result of [28].

Many of these structure results are motivated by the two-parameter family of minimal planar domains known as the Riemann examples mentioned earlier.

8.1. A few definitions. To set this up, we first recall a few properties of a complete properly embedded minimal planar domain $M \subset \mathbb{R}^3$.

First, it follows from a barrier argument of Meeks and Yau that one can find a stable embedded annulus between each pair of ends of M; a result of Fischer-Colbrie then implies that this stable surface has finite total curvature, so its ends are asymptotic to planes or half-catenoids. Since M is embedded, these planes or half-catenoids between its ends must all be parallel; this plane is the *limit tangent plane at infinity*.

Using these planes and half-catenoids in this way, Callahan, Frohman, Hoffman, and Meeks showed that the ends of M are ordered by height over this limit tangent plane at infinity. Moreover, a nice argument of Collin, Kusner, Meeks, and Rosenberg in [19] shows that there are at most two *limit ends*; these can only be on the "top" or on the "bottom".

Finally, since the coordinate functions are harmonic on any minimal surface, it follows from Stokes' theorem that the flux of a coordinate function x_i around a closed curve γ depends only on its homology class $[\gamma]$. Recall that the flux of x_i around γ is

$$\int_\gamma \frac{\partial x_i}{\partial n}, \tag{29}$$

where $\frac{\partial x_i}{\partial n}$ is the derivative of x_i in the conormal direction, i.e., the direction tangent to M but normal to γ. Using all three coordinates functions at once gives the *flux* map

$$\text{Flux}\colon [\gamma] \to \mathbb{R}^3. \tag{30}$$

8.2. The curvature estimate of [28]. We can now define $\mathcal{M}$ to be the space of properly embedded minimal planar domains in $\mathbb{R}^3$ with two limit ends, normalized so that every surface $M \in \mathcal{M}$ has horizontal limit tangent plane at infinity and the vertical component of its flux equals one. Here the horizontal flux is the projection of the flux vector to the limit tangent plane at infinity.

The main theorem of [28] is:

Theorem 8.1 (Theorem 5, [28]). *If a sequence $M_i \in \mathcal{M}$ has bounded horizontal flux, then the Gaussian curvature of the sequence is uniformly bounded.*

Remark 8.2. The main result in [29] is that there must always be two limit ends; thus this hypothesis can be removed from Theorem 8.1.

The first important application of the curvature estimates is to describe the geometry of a properly embedded minimal planar domain M with two limit ends:

> Such a surface M has bounded curvature and is conformally a compact Riemann surface punctured in a countable closed subset with two limit points; the spacing between consecutive ends is bounded from below in terms of the bound for the curvature; M is quasiperiodic in the sense that there exists a divergent sequence $V(n) \in \mathbb{R}^3$ such that the translated surfaces $M + V(n)$ converge to a properly embedded minimal surface of genus zero, two limit ends, a horizontal limit tangent plane at infinity and with the same flux as M.

Another particularly interesting consequence of Theorem 8.1 is a solution of an old conjecture of Nitsche:

Theorem 8.3 (Theorem 1, [28]). *Any complete minimal surface which is a union of simple closed curves in horizontal planes must be a catenoid.*

8.3. A heuristic argument for Theorem 8.1. Theorem 8.1 is best illustrated by considering the family of singly-periodic minimal surfaces known as the Riemann examples. After normalizing so the vertical flux is one and rotating so the horizontal flux points in the x_1-direction, there is a 1-parameter family parameterized by the length of the horizontal flux H. As $H \to 0$, this degenerates to a catenoid; when $H \to \infty$, this degenerates to a helicoid.

With this in mind, there is a simple heuristic argument for why such an estimate should hold. Namely, assume that $|A|^2(0) \to \infty$ for a sequence Σ_i and consider two possibilities:

1. The injectivity radius goes to zero.
2. Each Σ_i is uniformly simply connected.

In the first case, we would get short dividing curves in Σ_i; integrating around these would then imply that the flux was going to zero (violating the normalization).

In the second case, the results of Colding and the author in [8] give a limit plane through the origin which, by the uniqueness of the helicoid of Meeks and H. Rosenberg, is locally modelled by the helicoid for large i. As discussed above, this corresponds (roughly, at least) to a great deal of horizontal flux, violating the assumed bound on this.

The actual argument is much more complicated but has at least a similar flavor. One reason for the complications is that the dichotomy (1) or (2) above is more subtle; (1) involves the intrinsic distance while (2) uses the extrinsic distance. This dichotomy can now be made rigorous using the proof of the Calabi–Yau conjectures for embedded minimal surfaces with finite topology in [15]; however, the proof of Theorem 8.1 came before [15], so intrinsic and extrinsic distances were not yet known to be equivalent.

Appendix

A. Multi-valued graphs

We have used two notions of multi-valued graphs – namely, the one used in [4]–[7] and a generalization.

In [4]–[7], we defined multi-valued graphs as multi-sheeted covers of the punctured plane. To be precise, let D_r be the disk in the plane centered at the origin and of radius r and let $\mathcal{P}$ be the universal cover of the punctured plane $\mathbb{C} \setminus \{0\}$ with global polar coordinates (ρ, θ) so $\rho > 0$ and $\theta \in \mathbb{R}$. An N-valued graph of a function u on the annulus $D_s \setminus D_r$ is a single valued graph over

$$\{(\rho, \theta) \mid r \leq \rho \leq s,\ |\theta| \leq N\pi\}. \tag{31}$$

Note that the helicoid is the union of two infinite-valued graphs over the punctured plane together with the vertical axis.

Locally, the above multi-valued graphs give the complete picture for a ULSC sequence. However, the global picture can consist of several different multi-valued graphs glued together. To allow for this, we are forced to consider multi-valued graphs defined over the universal cover of $\mathbb{C} \setminus P$ where P is a discrete subset of the complex plane $\mathbb{C}$. We will see that the bound on the genus implies that P consists of at most two points.

B. The lamination theorem and one-sided curvature estimate

The first theorem that we recall shows that embedded minimal disks are either graphs or are part of double spiral staircases; moreover, a sequence of such disks with curvature blowing up converges to a foliation by parallel planes away from a singular

curve $\mathcal{S}$. This theorem is modelled on rescalings of the helicoid and the precise statement is as follows (we state the version for extrinsic balls; it was extended to intrinsic balls in [11]):

Theorem B.1 (Theorem 0.1 in [7]). *Let $\Sigma_i \subset B_{R_i} = B_{R_i}(0) \subset \mathbb{R}^3$ be a sequence of embedded minimal disks with $\partial\Sigma_i \subset \partial B_{R_i}$ where $R_i \to \infty$. If*

$$\sup_{B_1 \cap \Sigma_i} |A|^2 \to \infty, \tag{32}$$

then there exists a subsequence, Σ_j, and a Lipschitz curve $\mathcal{S}: \mathbb{R} \to \mathbb{R}^3$ such that after a rotation of $\mathbb{R}^3$:

1. *$x_3(\mathcal{S}(t)) = t$. (That is, $\mathcal{S}$ is a graph over the x_3-axis.)*
2. *Each Σ_j consists of exactly two multi-valued graphs away from $\mathcal{S}$ (which spiral together).*
3. *For each $1 > \alpha > 0$, $\Sigma_j \setminus \mathcal{S}$ converges in the C^α-topology to the foliation, $\mathcal{F} = \{x_3 = t\}_t$, of $\mathbb{R}^3$.*
4. *$\sup_{B_r(\mathcal{S}(t)) \cap \Sigma_j} |A|^2 \to \infty$ for all $r > 0$, $t \in \mathbb{R}$. (The curvatures blow up along $\mathcal{S}$),*

The second theorem that we need to recall asserts that every embedded minimal disk lying above a plane, and coming close to the plane near the origin, is a graph. Precisely this is the *intrinsic one-sided curvature estimate* which follows by combining [7] and [11]:

Theorem B.2. *There exists $\varepsilon > 0$, so that if*

$$\Sigma \subset \{x_3 > 0\} \subset \mathbb{R}^3 \tag{33}$$

is an embedded minimal disk with $\mathcal{B}_{2R}(x) \subset \Sigma \setminus \partial\Sigma$ and $|x| < \varepsilon R$, then

$$\sup_{\mathcal{B}_R(x)} |A_\Sigma|^2 \leq R^{-2}. \tag{34}$$

Theorem B.2 is in part used to prove the regularity of the singular set where the curvature is blowing up.

Note that the assumption in Theorem B.1 that the surfaces are disks is crucial and cannot even be replaced by assuming that the sequence is ULSC. To see this, observe that one can choose a one-parameter family of Riemann examples which is ULSC but where the singular set $\mathcal{S}$ is given by a *pair* of vertical lines. Likewise, the assumption in Theorem B.2 that Σ is simply connected is crucial as can be seen from the example of a rescaled catenoid, see (3). Under rescalings the catenoid converges (with multiplicity two) to the flat plane. Thus a neighborhood of the neck can be scaled arbitrarily close to a plane but the curvature along the neck becomes unbounded as it gets closer to the plane. Likewise, by considering the universal cover of the catenoid, one sees that embedded, and not just immersed, is needed in Theorem B.2.

C. The precise statement of the general compactness theorem

The precise statement of the compactness theorem for sequences that are neither necessarily ULSC nor with $\mathcal{S}_{\text{ulsc}} = \emptyset$ is the following:

Theorem C.1 ([8]). *Let $\Sigma_i \subset B_{R_i} = B_{R_i}(0) \subset \mathbb{R}^3$ be a sequence of compact embedded minimal planar domains with $\partial\Sigma_i \subset \partial B_{R_i}$ where $R_i \to \infty$. If*

$$\sup_{B_1 \cap \Sigma_i} |A|^2 \to \infty, \tag{35}$$

then there is a subsequence Σ_j, a closed set $\mathcal{S}$, and a lamination $\mathcal{L}'$ of $\mathbb{R}^3 \setminus \mathcal{S}$ so that:

(A) *For each $1 > \alpha > 0$, $\Sigma_j \setminus \mathcal{S}$ converges in the C^α-topology to the lamination $\mathcal{L}'$.*

(B) $\sup_{B_r(x)\cap\Sigma_j} |A|^2 \to \infty$ *as $j \to \infty$ for all $r > 0$ and $x \in \mathcal{S}$. (The curvatures blow up along $\mathcal{S}$).*

(C1) (C_{neck}) *from Theorem* 4.5 *holds for each point y in $\mathcal{S}_{\text{neck}}$.*

(C2) (C_{ulsc}) *from Theorem* 4.3 *holds locally near $\mathcal{S}_{\text{ulsc}}$. More precisely, each point y in $\mathcal{S}_{\text{ulsc}}$ comes with a sequence of multi-valued graphs in Σ_j that converge to the plane $\{x_3 = x_3(y)\}$. The convergence is in the C^∞ topology away from the point y and possibly also one other point in $\{x_3 = x_3(y)\} \cap \mathcal{S}_{\text{ulsc}}$. These two possibilities correspond to the two types of multi-valued graphs defined in Section A.*

(D) *The set $\mathcal{S}_{\text{ulsc}}$ is a union of Lipschitz curves transverse to the lamination. The leaves intersecting $\mathcal{S}_{\text{ulsc}}$ are planes foliating an open subset of $\mathbb{R}^3$ that does not intersect $\mathcal{S}_{\text{neck}}$. For the set $\mathcal{S}_{\text{neck}}$, we make no claim about the structure.*

(P) *Together* (C1) *and* (C2) *give a sequence of graphs or multi-valued graphs converging to a plane through each point of $\mathcal{S}$. If P is one of these planes, then each leaf of $\mathcal{L}'$ is either disjoint from P or is contained in P.*

Note that Theorem C.1 is a technical tool that is superseded by the stronger compactness theorems in the ULSC and non-ULSC cases, Theorem 4.3 and Theorem 4.5. This is because we will know by the no mixing theorem that either $\mathcal{S}_{\text{neck}} = \emptyset$ or $\mathcal{S}_{\text{ulsc}} = \emptyset$, so that these cover all possible cases.

References

[1] Choi, H. I., and Schoen, R., The space of minimal embeddings of a surface into a three-dimensional manifold of positive Ricci curvature. *Invent. Math.* **81** (1985), 387–394.

[2] Colding, T. H., and Minicozzi, W. P., II, Minimal surfaces. Courant Lecture Notes in Math. 4, New York University, Courant Institute of Mathematical Sciences, New York 1999.

[3] Colding, T. H., and Minicozzi, W. P., II, Estimates for parametric elliptic integrands. *Internat. Math. Res. Notices* **6** (2002), 291–297.

[4] Colding, T. H., and Minicozzi, W. P., II, The space of embedded minimal surfaces of fixed genus in a 3-manifold I; Estimates off the axis for disks. *Ann. of Math.* (2) **160** (2004), 27–68.

[5] Colding, T. H., and Minicozzi, W. P., II, The space of embedded minimal surfaces of fixed genus in a 3-manifold II; Multi-valued graphs in disks. *Ann. of Math.* (2) **160** (2004), 69–92.

[6] Colding, T. H., and Minicozzi, W. P., II, The space of embedded minimal surfaces of fixed genus in a 3-manifold III; Planar domains. *Ann. of Math.* (2) **160** (2004), 523–572.

[7] Colding, T. H., and Minicozzi, W. P., II, The space of embedded minimal surfaces of fixed genus in a 3-manifold IV; Locally simply connected. *Ann. of Math.* (2) **160** (2004), 573–615.

[8] Colding, T. H., and Minicozzi, W. P., II, The space of embedded minimal surfaces of fixed genus in a 3-manifold V: Fixed genus; math.DG/0509647.

[9] Colding, T. H., and Minicozzi, W. P., II, Multi-valued minimal graphs and properness of disks. *Internat. Math. Res. Notices* **21** (2002), 1111–1127.

[10] Colding, T. H., and Minicozzi, W. P., II, On the structure of embedded minimal annuli. *Internat. Math. Res. Notices* **29** (2002), 1539–1552.

[11] Colding, T. H., and Minicozzi, W. P., II, Disks that are double spiral staircases. *Notices Amer. Math. Soc.* **50** (3) (2003), 327–339.

[12] Colding, T. H., and Minicozzi, W. P., II, Embedded minimal disks: Proper versus nonproper – global versus local. *Trans. Amer. Math. Soc.* **356** (2004), 283–289.

[13] Colding, T. H., and Minicozzi, W. P., II, Complete properly embedded minimal surfaces in $\mathbb{R}^3$. *Duke Math. J.* **107** (2001), 421–426.

[14] Colding, T. H., and Minicozzi, W. P., II, Embedded minimal disks. In *Global theory of minimal surfaces*, Clay Math. Proc. 2, Amer. Math. Soc., Providence, RI, 2005, 405–438.

[15] Colding, T. H., and Minicozzi, W. P., II, The Calabi–Yau conjectures for embedded surfaces. Preprint; math.DG/0404197.

[16] Colding, T. H., and Minicozzi, W. P., II, Minimal submanifolds. *Bull. London Math. Soc.* **38** (3) (2006), 353–395.

[17] Colding, T. H., and Minicozzi, W. P., II, Shapes of embedded minimal surfaces. Preprint; math.DG/ 0511740.

[18] Collin, P., Topologie et courbure des surfaces minimales proprement plongées de $\mathbb{R}^3$. *Ann. of Math.* (2) **145** (1997), 1–31.

[19] Collin, P., Kusner, R., Meeks, W. H., III, and Rosenberg, H., The topology, geometry and conformal structure of properly embedded minimal surfaces. *J. Differential Geom.* **67** (2) (2004), 377–393.

[20] Dean, B., Embedded minimal disks with prescribed curvature blowup. *Proc. Amer. Math. Soc.* **134** (2006), 1197–1204.

[21] Hoffman, D., Weber, M., and Wolf, M., An embedded genus-one helicoid. math.DG/ 0401080.

[22] Hoffman, D., Weber, M., and Wolf, M., An embedded genus-one helicoid. *Proc. Nat. Acad. Sci. U.S.A.* **102** (46) (2005), 16566–16568.

[23] Kapouleas, N., Constructions of minimal surfaces by gluing minimal immersions. In *Global theory of minimal surfaces*, Clay Math. Proc. 2, Amer. Math. Soc., Providence, RI, 2005, 489–524.

[24] Martin, F., Complete nonorientable minimal surfaces in $\mathbb{R}^3$. In *Global theory of minimal surfaces*, Clay Math. Proc. 2, Amer. Math. Soc., Providence, RI, 2005, 371–380.

[25] Meeks, W. H., III, The regularity of the singular set in the Colding and Minicozzi lamination theorem. *Duke Math. J.* **123** (2) (2004), 329–334.

[26] Meeks, W. H., III, The lamination metric for the Colding-Minicozzi minimal lamination. *Illinois J. Math.* **49** (2) (2005), 645–658.

[27] Meeks, W. H., III, and Perez, J., Conformal properties in classical minimal surface theory. In *Surveys in Diff. Geom. IX: Eigenvalues of Laplacians and other geometric operators* (ed. by A. Grigor'yan and S. T. Yau), International Press, Somerville, MA, 2004, 275–335.

[28] Meeks, W. H., III, Perez, J., and Ros, A., The geometry of minimal surfaces of finite genus I; Curvature estimates and quasiperiodicity. *J. Differential Geom.* **66** (2004), 1–45.

[29] Meeks, W. H., III, Perez, J., and Ros, A., The geometry of minimal surfaces of finite genus II; Nonexistence of one limit end examples. *Invent. Math.* **158** (2) (2004), 323–341.

[30] Meeks, W. H., III, Perez, J., and Ros, A., The geometry of minimal surfaces of finite genus III; bounds on the topology and index of classical minimal surfaces. Preprint.

[31] Meeks, W. H., III, and Rosenberg, H., The uniqueness of the helicoid and the asymptotic geometry of properly embedded minimal surfaces with finite topology. *Ann. of Math.* (2) **161** (2) (2005), 727–758.

[32] Meeks, W. H., III, and Weber, M., Existence of bent helicoids and the regularity of the singular set in the Colding-Minicozzi lamination theorem. In preparation.

[33] Meeks, W. H., III, and Yau, S. T., The classical Plateau problem and the topology of three dimensional manifolds. *Topology* **21** (1982), 409–442.

[34] Meeks, W. H., III, and Yau, S. T., Topology of three-dimensional manifolds and the embedding problems in minimal surface theory. *Ann. of Math.* (2) **112** (3) (1980), 441–484.

[35] J. Perez, J., Limits by rescalings of minimal surfaces: Minimal laminations, curvature decay and local pictures. Notes for the workshop "Moduli Spaces of Properly Embedded Minimal Surfaces", American Institute of Mathematics, Palo Alto, California, 2005.

[36] Rosenberg, H., Some recent developments in the theory of minimal surfaces in 3-manifolds. IMPA Mathematical Publications, 24th Brazilian Mathematics Colloquium, Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 2003.

[37] Schoen, R., Estimates for stable minimal surfaces in three–dimensional manifolds. In *Seminar on Minimal Submanifolds*, Ann. of Math. Stud. 103, Princeton University Press, Princeton, N.J., 1983, 111–126.

[38] Traizet, M., Construction of minimal surfaces by gluing Weierstrass representations. In *Global theory of minimal surfaces*, 439–452, Clay Math. Proc. 2, Amer. Math. Soc., Providence, RI, 2005.

Department of Mathematics, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, U.S.A.

E-mail: minicozz@jhu.edu

Floer homology in symplectic geometry and in mirror symmetry

Yong-Geun Oh* and Kenji Fukaya

Abstract. In this article the authors review what the Floer homology is and what it does in symplectic geometry both in the closed string and in the open string context. In the first case, the authors will explain how the chain level Floer theory leads to the C^0 symplectic invariants of Hamiltonian flows and to the study of topological Hamiltonian dynamics. In the second case, the authors explain how Floer's original construction of Lagrangian intersection Floer homology is obstructed in general as soon as one leaves the category of exact Lagrangian submanifolds. They will survey the construction of the Floer complex and describe its obstruction in terms of the filtered A_∞-algebras. This can be promoted to the level of A_∞-category (Fukaya category) of symplectic manifolds. Some applications of this general machinery to the study of the topology of Lagrangian embeddings in relation to symplectic topology and to mirror symmetry are also reviewed.

Mathematics Subject Classification (2000). Primary 53D40; Secondary 14J32.

Keywords. Floer homology, Hamiltonian flows, Lagrangian submanifolds, A_∞-structure, mirror symmetry.

1. Prologue

The Darboux theorem in symplectic geometry manifests *flexibility* of the group of symplectic transformations. On the other hand, the following celebrated theorem of Eliashberg [El1] revealed subtle *rigidity* of symplectic transformations: *The subgroup* $\mathrm{Symp}(M, \omega)$ *consisting of symplectomorphisms is closed in* $\mathrm{Diff}(M)$ *with respect to the* C^0*-topology.*

This demonstrates that the study of symplectic topology is subtle and interesting. Eliashberg's theorem relies on a version of non-squeezing theorem as proven by Gromov [Gr]. Gromov [Gr] uses the machinery of pseudo-holomorphic curves to prove his theorem. There is also a different proof by Ekeland and Hofer [EH] of the classical variational approach to Hamiltonian systems. The interplay between these two facets of symplectic geometry has been the main locomotive in the development of symplectic topology since Floer's pioneering work on his 'semi-infinite' dimensional homology theory, now called the *Floer homology theory*.

*Oh thanks A. Weinstein and late A. Floer for putting everlasting marks on his mathematics. Both authors thank H. Ohta and K. Ono for a fruitful collaboration on the Lagrangian intersection Floer theory where some part of this survey is based on.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

As in classical mechanics, there are two most important boundary conditions in relation to Hamilton's equation $\dot{x} = X_H(t, x)$ on a general symplectic manifold: one is the *periodic* boundary condition $\gamma(0) = \gamma(1)$, and the other is the *Lagrangian* boundary condition $\gamma(0) \in L_0,\ \gamma(1) \in L_1$ for a given pair (L_0, L_1) of two Lagrangian submanifolds: A submanifold $i : L \hookrightarrow (M, \omega)$ is called Lagrangian if $i^*\omega = 0$ and $\dim L = \frac{1}{2} \dim M$. The latter replaces the *two-point* boundary condition in classical mechanics.

In either of the above two boundary conditions, we have a version of the *least action principle*: a solution of Hamilton's equation corresponds to a critical point of the action functional on a suitable path space with the corresponding boundary condition. For the periodic boundary condition we consider the free loop space

$$\mathcal{L}M = \{\gamma : S^1 \to M\},$$

and for the Lagrangian boundary condition we consider the space of paths connecting

$$\Omega(L_0, L_1) = \{\gamma : [0, 1] \to M \mid \gamma(0) \in L_0,\ \gamma(1) \in L_1\}.$$

Both $\mathcal{L}M$ and $\Omega(L_0, L_1)$ have countable number of connected components. For the case of $\mathcal{L}M$, it has a distinguished component consisting of the contractible loops. On the other hand, for the case of $\Omega(L_0, L_1)$ there is no such distinguished component in general.

Daunting Questions. For a given time dependent Hamiltonian $H = H(t, x)$ on (M, ω), does there exist a solution of the Hamilton equation $\dot{x} = X_H(t, x)$ with the corresponding boundary conditions? If so, how many different ones can exist?

One crucial tool for the study of these questions is the least action principle. Another seemingly trivial but crucial observation is that when $H \equiv 0$ for the closed case and when $L_1 = L_0$ (and $H \equiv 0$) for the open case, there are "many" solutions given by *constant* solutions. It turns out that these two ingredients, combined with Gromov's machinery of pseudo-holomorphic curves, can be utilized to study each of the above questions, culminating in Floer's proof of Arnold's conjecture for the fixed points [Fl2], and for the intersection points of L with its Hamiltonian deformation $\phi_H^1(L)$ [Fl1] for the exact case respectively.

We divide the rest of our exposition into two categories, one in the closed string and the other in the open string context.

2. Floer theory of Hamiltonian fixed points

2.1. Construction. On a symplectic manifold (M, ω), for each given time-periodic Hamiltonian H i.e., H with $H(t, x) = H(t + 1, x)$, there exists an analog $\mathcal{A}_H$ to the classical action functional defined on a suitable covering space of the free loop space.

To exploit the fact that in the vacuum, i.e., when $H \equiv 0$, we have many constant solutions all lying in the distinguished component of the free loop space $\mathcal{L}(M)$,

$$\mathcal{L}_0(M) = \{\gamma\colon [0,1] \to M \mid \gamma(0) = \gamma(1),\ \gamma \text{ contractible}\}$$

one studies the contractible periodic solutions and so the action functional on $\mathcal{L}_0(M)$. The covering space, denoted by $\widetilde{\mathcal{L}}_0(M)$, is realized by the set of suitable equivalence classes $[z, w]$ of the pair (z, w) where $z\colon S^1 \to M$ is a loop and $w\colon D^2 \to M$ is a disc bounding z. Then $\mathcal{A}_H$ is defined by

$$\mathcal{A}_H([z,w]) = -\int w^*\omega - \int_0^1 H(t,\gamma(t))\,dt. \tag{2.1}$$

This reduces to the classical action $\int p\,dq - H\,dt$ if we define the canonical symplectic form as $\omega_0 = \sum_j dq^j \wedge dp_j$ on the phase space $\mathbb{R}^{2n} \cong T^*\mathbb{R}^n$.

To do Morse theory, one needs to introduce a metric on $\Omega(M)$, which is done by introducing an almost complex structure J that is compatible to ω (in that the bilinear form $g_J := \omega(\cdot, J\cdot)$ defines a Riemannian metric on M) and integrating the norm of the tangent vectors of the loop γ. To make the Floer theory a more flexible tool to use, one should allow this J to be time-dependent.

A computation shows that the negative L^2-gradient flow equation of the action functional for the path $u\colon \mathbb{R} \times S^1 \to M$ is the following *nonlinear* first order partial differential equation

$$\frac{\partial u}{\partial \tau} + J\left(\frac{\partial u}{\partial t} - X_H(t,u)\right) = 0. \tag{2.2}$$

The rest points of this gradient flow are the periodic orbits of $\dot{x} = X_H(t,x)$. Note that when $H = 0$, this equation becomes the pseudo-holomorphic equation

$$\overline{\partial}_J(u) = \frac{\partial u}{\partial \tau} + J\frac{\partial u}{\partial t} = 0 \tag{2.3}$$

which has many constant solutions. Following Floer [Fl2], for each given *nondegenerate* H, i.e., one whose time-one map ϕ_H^1 has the linearization with no eigenvalue 1, we consider a vector space $CF(H)$ consisting of *Novikov Floer chains*

Definition 2.1. For each formal sum

$$\beta = \sum_{[z,w] \in \mathrm{Crit}\mathcal{A}_H} a_{[z,w]}[z,w],\ a_{[z,w]} \in \mathbb{Q} \tag{2.4}$$

we define the *support* of β by the set

$$\operatorname{supp}\beta = \{[z,w] \in \operatorname{Crit}\mathcal{A}_H \mid a_{[z,w]} \neq 0 \text{ in } (2.4)\}.$$

We call β a Novikov Floer chain or (simply a Floer chain) if it satisfies the condition

$$\#\{[z,w] \in \operatorname{supp}\beta \mid \mathcal{A}_H([z,w]) \geq \lambda\} < \infty$$

for all $\lambda \in \mathbb{R}$ and define $CF(H)$ to be the set of Novikov Floer chains.

$CF(H)$ can be considered either as a $\mathbb{Q}$-vector space or a module over the Novikov ring Λ_ω of (M, ω). Each Floer chain β as a $\mathbb{Q}$-chain can be regarded as the union of "unstable manifolds" of the generators $[z, w]$ of β, which has a 'peak'. There is the natural Floer boundary map $\partial = \partial_{(H,J)}\colon CF(H) \to CF(H)$ i.e., a linear map satisfying $\partial\partial = 0$. The pair $(CF(H), \partial_{(H,J)})$ is the *Floer complex* and the quotient

$$HF_*(H, J; M) := \ker \partial_{(H,J)} / \operatorname{im} \partial_{(H,J)}$$

is the *Floer homology*. By now the general construction of this Floer homology has been carried out by Fukaya–Ono [FOn], Liu–Tian [LT1], and Ruan [Ru] in its complete generality, after the construction had been previously carried out by Floer [Fl2], Hofer–Salamon [HS] and by Ono [On] in some special cases.

The Floer homology $HF_*(H, J; M)$ also has the ring structure arising from the pants product, which becomes the *quantum product* on $H^*(M)$ in "vacuum", i.e. when $H \equiv 0$. The module $H^*(M) \otimes \Lambda_\omega$ with this ring structure is the *quantum cohomology ring* denoted by $QH^*(M)$. We denote by $a \cdot b$ the quantum product of two quantum cohomology classes a and b.

2.2. Spectral invariants and spectral norm. Knowing the existence of periodic orbits of a given Hamiltonian flow, the next task is to organize the collection of the actions of different periodic orbits and to study their relationships.

We first collect the actions of all possible periodic orbits, *including their quantum contributions*, and define the *action spectrum* of H by

$$\operatorname{Spec}(H) := \{\mathcal{A}_H([z, w]) \in \mathbb{R} \mid [z, w] \in \widetilde{\Omega}_0(M),\ d\mathcal{A}_H([z, w]) = 0\} \tag{2.5}$$

i.e., the set of critical values of $\mathcal{A}_H\colon \widetilde{\mathcal{L}}_0(M) \to \mathbb{R}$. In general this set is a countable subset of $\mathbb{R}$ on which the (spherical) period group Γ_ω acts. Motivated by classical Morse theory and mini-max theory, one would like to consider a sub-collection of critical values that are *homologically essential*: each non-trivial cohomology class gives rise to a mini-max value, which cannot be pushed further down by the gradient flow. One crucial ingredient in the classical mini-max theory is a choice of semi-infinite cycles that are linked under the gradient flow.

Applying this idea in the context of chain level Floer theory, Oh generalized his previous construction [Oh4], [Oh5] to the non-exact case in [Oh8], [Oh10]. We define the *level* of a Floer chain β by the maximum value

$$\lambda_H(\beta) := \max_{[z,w]} \{\mathcal{A}_H([z, w]) \mid [z, w] \in \operatorname{supp} \beta\}. \tag{2.6}$$

Now for each $a \in QH^k(M)$ and a generic J, Oh considers the mini-max values

$$\rho(H, J; a) = \inf_\alpha \{\lambda_H(\alpha) \mid \alpha \in CF_{n-k}(H),\ \partial_{(H,J)}\alpha = 0,\ [\alpha] = a^\flat\} \tag{2.7}$$

where $2n = \dim M$ and proves that this number is independent of J [Oh8]. The common number denoted by $\rho(H; a)$ is called the *spectral invariant* associated to

the Hamiltonian H relative to the class $a \in QH^*(M)$. The collection of the values $\rho(H; a)$ extend to arbitrary smooth Hamiltonian function H, *whether H is nondegenerate or not*, and satisfy the following basic properties.

Theorem 2.2 ([Oh8], [Oh10]). *Let (M, ω) be an arbitrary closed symplectic manifold. For any given quantum cohomology class $0 \neq a \in QH^*(M)$, we have a continuous function denoted by $\rho = \rho(H; a) : C^\infty_m([0, 1]\times M)\times(QH^*(M)\setminus\{0\}) \to \mathbb{R}$ which satisfies the following axioms. Let $H,\ F \in C^\infty_m([0, 1] \times M)$ be smooth Hamiltonian functions and $a \neq 0 \in QH^*(M)$. Then we have:*

1. (Projective invariance) $\rho(H; \lambda a) = \rho(H; a)$ *for any* $0 \neq \lambda \in \mathbb{Q}$.
2. (Normalization) *For a quantum cohomology class a, we have $\rho(\underline{0}; a) = v(a)$ where $\underline{0}$ is the zero function and $v(a)$ is the valuation of a on $QH^*(M)$.*
3. (Symplectic invariance) $\rho(\eta^* H; a) = \rho(H; a)$ *for any* $\eta \in \mathrm{Symp}(M, \omega)$.
4. (Homotopy invariance) *For any $H,\ K$ with $[H] = [K]$, $\rho(H; a) = \rho(K; a)$.*
5. (Multiplicative triangle inequality) $\rho(H\#F; a \cdot b) \leq \rho(H; a) + \rho(F; b)$.
6. (C^0-continuity) $|\rho(H; a) - \rho(F; a)| \leq \|H - F\|$. *In particular, the function $\rho_a : H \mapsto \rho(H; a)$ is C^0-continuous.*
7. (Additive triangle inequality) $\rho(H; a + b) \leq \max\{\rho(H; a), \rho(H; b)\}$.

Under the canonical one-one correspondence between (*smooth*) H (satisfying $\int_M H_t = 0$) and its Hamiltonian path $\phi_H : t \mapsto \phi^t_H$, we denote by $[H]$ the path-homotopy class of the Hamiltonian path $\phi_H : [0, 1] \to \mathrm{Ham}(M, \omega)$; $\phi_H(t) = \phi^t_H$ with fixed end points, and by $\widetilde{\mathrm{Ham}}(M, \omega)$ the set of $[H]$'s which represents the universal covering space of $\mathrm{Ham}(M, \omega)$.

This theorem generalizes the results on the exact case by Viterbo [V2], Oh [Oh4], [Oh5] and Schwarz [Sc] to the non-exact case. The axioms 1 and 7 already hold at the level of *cycles* or for λ_H and follow immediately from its definition. All other axioms are proved in [Oh8] except the homotopy invariance for the *irrational* symplectic manifolds which is proven in [Oh10]. The additive triangle inequality was explicitly used by Entov and Polterovich in their construction of some quasi-morphisms on $\mathrm{Ham}(M, \omega)$ [EnP]. The axiom of homotopy invariance implies that $\rho(\,\cdot; a)$ projects down to $\widetilde{\mathrm{Ham}}(M, \omega)$. It is a consequence of the following spectrality axiom, which is proved for *any* H on *rational* (M, ω) in [Oh8] and just for *nondegenerate* H on *irrational* (M, ω) [Oh10]:

8. (Nondegenerate spectrality) *For nondegenerate H, the mini-max values $\rho(H; a)$ lie in* $\mathrm{Spec}(H)$, *i.e. they are critical values of $\mathcal{A}_H$ for all $a \in QH^*(M) \setminus \{0\}$.*

The following is still an open problem.

Question 2.3. Let (M, ω) be a irrational symplectic manifold, i.e., the period group $\Gamma_\omega = \{\omega(A) \mid A \in \pi_2(M)\}$ be a dense subgroup of $\mathbb{R}$. Does $\rho(H; a)$ still lie in $\mathrm{Spec}(H)$ for all $a \neq 0$ for *degenerate* Hamiltonian H?

It turns out that the invariant $\rho(H; 1)$ can be used to construct a canonical invariant norm on $\mathrm{Ham}(M, \omega)$ of the Viterbo type which is called the *spectral norm*. To describe this construction, we start by reviewing the definition of the Hofer norm $\|\phi\|$ of a Hamiltonian diffeomorphism ϕ.

There are two natural operations on the space of Hamiltonians H: one the *inverse* $H \mapsto \bar{H}$ where $\bar{H}$ is the Hamiltonian generating the inverse flow $(\phi_H^t)^{-1}$ and the *product* $(H, F) \mapsto H \# F$ where $H \# F$ is the one generating the composition flow $\phi_H^t \circ \phi_F^t$. Hofer [H] introduced an invariant norm on $\mathrm{Ham}(M, \omega)$. Hofer also considered its $L^{(1,\infty)}$-version $\|\phi\|$ defined by

$$\|\phi\| = \inf_{H \mapsto \phi} \|H\|; \quad \|H\| = \int_0^1 (\max H_t - \min H_t)\, dt$$

where $H \mapsto \phi$ stands for $\phi = \phi_H^1$. We call $\|H\|$ the $L^{(1,\infty)}$*-norm* of H and $\|\phi\|$ the $L^{(1,\infty)}$ *Hofer norm* of ϕ.

Making use of the spectral invariant $\rho(H; 1)$, Oh defined in [Oh9] a function $\gamma : C_m^\infty([0, 1] \times M) \to \mathbb{R}$ by

$$\gamma(H) = \rho(H; 1) + \rho(\bar{H}; 1)$$

on $C_m^\infty([0, 1] \times M)$, whose definition is more *topological* than $\|H\|$. For example, γ canonically projects down to a function on $\widetilde{\mathrm{Ham}}(M, \omega)$ by the homotopy invariance axiom while $\|H\|$ does not. Obviously $\gamma(H) = \gamma(\bar{H})$. The inequality $\gamma(H) \leq \|H\|$ was also shown in [Oh4], [Oh9] and the inequality $\gamma(H) \geq 0$ follows from the triangle inequality applied to $a = b = 1$ and from the normalization axiom $\rho(\underline{0}; 1) = 0$.

Now we define a non-negative function $\gamma : \mathrm{Ham}(M, \omega) \to \mathbb{R}_+$ by $\gamma(\phi) := \inf_{H \mapsto \phi} \gamma(H)$. Then the following theorem is proved in [Oh9].

Theorem 2.4 ([Oh9]). *Let (M, ω) be any closed symplectic manifold. Then*

$$\gamma : \mathrm{Ham}(M, \omega) \to \mathbb{R}_+$$

defines a (*non-degenerate*) *norm on* $\mathrm{Ham}(M, \omega)$ *which satisfies the following additional properties:*

1. $\gamma(\eta^{-1}\phi\eta) = \gamma(\phi)$ *for any symplectic diffeomorphism* η*;*
2. $\gamma(\phi^{-1}) = \gamma(\phi)$, $\gamma(\phi) \leq \|\phi\|$.

Oh then applied the function $\gamma = \gamma(H)$ to the study of the geodesic property of Hamiltonian flows [Oh7], [Oh9].

Another interesting application of spectral invariants is a new construction of *quasi-morphisms* on $\mathrm{Ham}(M,\omega)$ carried out by Entov and Polterovich [EnP]. Recall that for a closed (M,ω), there exists no non-trivial homomorphism to $\mathbb{R}$ because $\mathrm{Ham}(M,\omega)$ is a simple group [Ba]. However for a certain class of *semi-simple* symplectic manifolds, e.g. for (S^2,ω), $(S^2\times S^2, \omega\oplus\omega)$, $(\mathbb{C}P^n,\omega_{FS})$, Entov and Polterovich [EnP] produced non-trivial quasi-morphisms, exploiting the spectral invariants $\rho(e,\cdot)$ corresponding to a certain idempotent element e of the quantum cohomology ring $QH^*(M)$.

It would be an important problem to unravel what the true meaning of Gromov's pseudo-holomorphic curves or of the Floer homology in general is in regard to the underlying symplectic structure.

3. Towards topological Hamiltonian dynamics

We note that the construction of spectral invariants largely depends on the smoothness (or at least differentiability) of Hamiltonians H because it involves the study of Hamilton's equation $\dot{x} = X_H(t,x)$. If H is smooth there is a one-one correspondence between H and its flow ϕ_H^t. However this correspondence breaks down when H does not have enough regularity, e.g., if H is only continuous or even C^1 *because the fundamental existence and uniqueness theorems of ODE's fail.*

However the final outcome $\rho(H;a)$ still makes sense for and can be extended to a certain natural class of C^0-functions H. Now a natural questions to ask is:

Question 3.1. Can we define the notion of topological Hamiltonian dynamical systems? If so, what is the dynamical meaning of the numbers $\rho(H;a)$ when H is just continuous but not differentiable?

These questions led to the notions of *topological Hamiltonian paths* and *Hamiltonian homeomorphisms* in [OM].

Definition 3.2. A continuous path $\lambda\colon [0,1]\to \mathrm{Homeo}(M)$ with $\lambda(0)=\mathrm{id}$ is called a topological Hamiltonian path if there exists a sequence of smooth Hamiltonians $H_i\colon [0,1]\times M\to\mathbb{R}$ such that

1. H_i converges in the $L^{(1,\infty)}$-topology (or Hofer topology) of Hamiltonians, and
2. $\phi_{H_i}^t\to\lambda(t)$ uniformly converges on $[0,1]$.

We say that the $L^{(1,\infty)}$-limit of any such sequence H_i is a *Hamiltonian* of the topological Hamiltonian flow λ. The following uniqueness result is proved in [Oh12].

Theorem 3.3 ([Oh12]). *Let λ be a topological Hamiltonian path. Suppose that there exist two sequences H_i and H_i' satisfying the conditions in Definition* 3.2. *Then their limits coincide as an $L^{(1,\infty)}$-function.*

The proof of this theorem is a modification of Viterbo's proof [V3] of a similar uniqueness result for the C^0 Hamiltonians, combined with a structure theorem of topological Hamiltonians which is also proven in [Oh12]. An immediate corollary is the following extension of the spectral invariants to the space of topological Hamiltonian paths.

Definition 3.4. Suppose λ is a topological Hamiltonian path and let H_i be the sequence of smooth Hamiltonians that converges in $L^{(1,\infty)}$-topology and whose associated Hamiltonian path ϕ_{H_i} converges to λ uniformly. We define

$$\rho(\lambda; a) = \lim_{i\to\infty} \rho(H_i; a).$$

The uniqueness theorem of topological Hamiltonians and the $L^{(1,\infty)}$ continuity property of ρ imply that this definition is well defined.

Definition 3.5. A homeomorphism h of M is a Hamiltonian homeomorphism if there exists a sequence of smooth Hamiltonians $H_i : [0, 1] \times M \to \mathbb{R}$ such that

1. H_i converges in the $L^{(1,\infty)}$-topology of Hamiltonians, and
2. the Hamiltonian path $\phi_{H_i} : t \mapsto \phi^t_{H_i}$ uniformly converges on $[0, 1]$ in the C^0-topology of Homeo(M), and $\phi^1_{H_i} \to h$.

We denote by Hameo(M, ω) the set of such homeomorphisms.

Motivated by Eliashberg's rigidity theorem we also define the group Sympeo(M, ω) as the subgroup of Homeo(M) consisting of the C^0-limits of symplectic diffeomorphisms. Then Oh and Müller [OM] proved the following theorem.

Theorem 3.6 ([OM]). Hameo(M, ω) *is a path-connected normal subgroup of* Sympeo$_0(M, \omega)$*, the identity component of* Sympeo(M, ω).

One can easily derive that Hameo(M, ω) is a proper subgroup of Sympeo$_0(M, \omega)$ whenever the so-called mass flow homomorphism [Fa] is non-trivial or there exists a symplectic diffeomorphism that has no fixed point, e.g., T^{2n} [OM]. In fact, we conjecture that this is always the case.

Conjecture 3.7. The group Hameo(M, ω) is a proper subgroup of Sympeo$_0(M, \omega)$ for any closed symplectic manifold (M, ω).

A case of particular interest is the case $(M, \omega) = (S^2, \omega)$. In this case, together with the smoothing result proven in [Oh11], the affirmative answer to this conjecture would yield a negative response to the following open question in area preserving dynamical systems. See [Fa] for the basic theorems on the measure preserving homeomorphisms in dimension greater than or equal to 3.

Question 3.8. Is the identity component of the group of area preserving homeomorphisms on S^2 a simple group?

4. Floer theory of Lagrangian intersections

Floer's original definition [Fl1] of the homology $HF(L_0, L_1)$ of Lagrangian submanifolds meets many obstacles when one attempts to generalize his definition beyond the exact cases, i.e. the case

$$L_0 = L, \quad L_1 = \phi(L) \quad \text{with } \pi_2(M, L) = \{0\}.$$

In this exposition we will consider the cases of Lagrangian submanifolds that are not necessarily exact. In the open string case of Lagrangian submanifolds one has to deal with the phenomenon of bubbling-off discs besides bubbling-off spheres. One crucial difference between the former and the latter is that the former is a phenomenon of codimension one while the latter is of codimension two. This difference makes the general Lagrangian intersection Floer theory show a very different perspective compared to the Floer theory of Hamiltonian fixed points. For example, for the intersection case in general, one has to study the theory *in the chain level*, which forces one to consider the chain complexes themselves. Then the meaning of invariance of the resulting objects is much more non-trivial to define compared to that of Gromov–Witten invariants for which one can work in the level of homology.

There is one particular case that Oh singled out in [Oh1] for which the original version of Floer cohomology is well defined and invariant just under the change of almost complex structures and under the Hamiltonian isotopy. This is the case of *monotone* Lagrangian submanifolds with *minimal Maslov number* $\Sigma_L \geq 3$:

Definition 4.1. A Lagrangian submanifold $L \subset (M, \omega)$ is *monotone* if there exists a constant $\lambda \geq 0$ such that $\omega(A) = \lambda\mu(A)$ for all elements $A \in \pi_2(M, L)$. The minimal Maslov number is defined by the integer

$$\Sigma_L = \min\{\mu(\beta) \mid \beta \in \pi_2(M, L),\ \mu(\beta) > 0\}.$$

We will postpone further discussion on this particular case until later in this survey but proceed with describing the general story now.

To obtain the maximal possible generalization of Floer's construction, it is crucial to develop a proper *off-shell* formulation of the relevant Floer moduli spaces.

4.1. Off-shell formulation. We consider the space of paths

$$\Omega = \Omega(L_0, L_1) = \{\ell : [0, 1] \to P \mid \ell(0) \in L_0, \ell(1) \in L_1\}.$$

On this space we are given the *action one-form* α defined by

$$\alpha(\ell)(\xi) = \int_0^1 \omega(\dot{\ell}(t), \xi(t))\, dt$$

for each tangent vector $\xi \in T_\ell\Omega$. From this expression it follows that

$$\text{Zero}(\alpha) = \{\ell_p : [0, 1] \to M \mid p \in L_0 \cap L_1, \quad \ell_p \equiv p\}.$$

Using the Lagrangian property of (L_0, L_1), a straightforward calculation shows that this form is *closed*. Note that $\Omega(L_0, L_1)$ is not connected but has countably many connected components. We will work on a particular fixed connected component of $\Omega(L_0, L_1)$. We pick up a based path $\ell_0 \in \Omega(L_0, L_1)$ and consider the corresponding component $\Omega(L_0, L_1; \ell_0)$, and then define a covering space

$$\pi : \widetilde{\Omega}(L_0, L_1; \ell_0) \to \Omega(L_0, L_1; \ell_0)$$

on which we have a single valued action functional such that $d\mathcal{A} = -\pi^*\alpha$. One can repeat Floer's construction similarly as in the closed case replacing $\mathcal{L}_0(M)$ by the chosen component of the path space $\Omega(L_0, L_1)$. We refer to [FOOO] for the details of this construction. We then denote by $\Pi(L_0, L_1; \ell_0)$ the group of deck transformations. We define the associated Novikov ring $\Lambda(L_0, L_1; \ell_0)$ as a completion of the group ring $\mathbb{Q}[\Pi(L_0, L_1; \ell_0)]$.

Definition 4.2. $\Lambda_k(L_0, L_1; \ell_0)$ denotes the set of all (infinite) sums

$$\sum_{\substack{g \in \Pi(L_0, L_1; \ell_0) \\ \mu(g) = k}} a_g [g]$$

with $a_g \in \mathbb{Q}$ and such that for each $C \in \mathbb{R}$,

$$\#\{g \in \Pi(L_0, L_1; \ell_0) \mid E(g) \leq C,\ a_g \neq 0\} < \infty.$$

We put $\Lambda(L_0, L_1; \ell_0) = \bigoplus_k \Lambda_k(L_0, L_1; \ell_0)$.

We call this graded ring the *Novikov ring* of the pair (L_0, L_1) relative to the path ℓ_0. Note that this ring depends on L and ℓ_0. In relation to mirror symmetry one needs to consider a family of Lagrangian submanifolds and to use a universal form of this ring. The following ring was introduced in [FOOO], which plays an important role in the rigorous formulation of homological mirror symmetry conjecture.

Definition 4.3 (Universal Novikov ring). We define

$$\Lambda_{\text{nov}} = \Big\{ \sum_{i=1}^{\infty} a_i T^{\lambda_i} \ \Big|\ a_i \in \mathbb{Q},\ \lambda_i \in \mathbb{R},\ \lambda_i \leq \lambda_{i+1},\ \lim_{i \to \infty} \lambda_i = \infty \Big\}, \tag{4.1}$$

$$\Lambda_{0,\text{nov}} = \Big\{ \sum_{i=1}^{\infty} a_i T^{\lambda_i} \in \Lambda_{\text{nov}} \ \Big|\ \lambda_i \geq 0 \Big\}. \tag{4.2}$$

In the above definitions of Novikov rings, one can replace $\mathbb{Q}$ by other commutative rings with unit, e.g. $\mathbb{Z}$, $\mathbb{Z}_2$ or $\mathbb{Q}[e]$ with a formal variable e.

There is a natural filtration on these rings by the valuation $v\colon \Lambda_{\text{nov}},\ \Lambda_{0,\text{nov}} \to \mathbb{R}$ defined by

$$v\Big(\sum_{i=1}^{\infty} a_i T^{\lambda_i} \Big) := \lambda_1. \tag{4.3}$$

This is well defined by the definition of the Novikov ring and induces a filtration $F^\lambda\Lambda_{\text{nov}} := v^{-1}([\lambda, \infty))$ on Λ_{nov}. The function $e^{-v}\colon \Lambda_{\text{nov}} \to \mathbb{R}_+$ also provides a natural non-Archimedean norm on Λ_{nov}. We call the induced topology on Λ_{nov} a *non-Archimedean topology*.

We now assume that L_0 intersects L_1 transversely and form the $\mathbb{Q}$-vector space $CF(L_0, L_1)$ over the set $\text{span}_{\mathbb{Q}}\,\text{Crit}\,\mathcal{A}$ similarly as $CF(H)$. Now let $p,\ q \in L_0 \cap L_1$. We denote by $\pi_2(p, q) = \pi_2(p, q; L_0, L_1)$ the set of homotopy classes of smooth maps $u\colon [0, 1] \times [0, 1] \to M$ relative to the boundary

$$u(0, t) \equiv p, \quad u(1, t) = q; \quad u(s, 0) \in L_0, \quad u(s, 1) \in L_1,$$

by $[u] \in \pi_2(p, q)$ the homotopy class of u, and by B a general element in $\pi_2(p, q)$. For given $B \in \pi_2(p, q)$ we denote by $\text{Map}(p, q; B)$ the set of such w's in the class B. Each element $B \in \pi_2(p, q)$ induces a map given by the obvious gluing map $[p, w] \mapsto [q, w \# u]$ for $u \in \text{Map}(p, q; B)$. There is also the natural gluing map

$$\pi_2(p, q) \times \pi_2(q, r) \to \pi_2(p, r)$$

induced by the concatenation $(u_1, u_2) \mapsto u_1 \# u_2$.

4.2. Floer moduli spaces and Floer operators. Now for each given $J = \{J_t\}_{0\le t\le 1}$ and $B \in \pi_2(p, q)$ we define the moduli space $\widetilde{\mathcal{M}}(p, q; B)$ consisting of finite energy solutions of the Cauchy–Riemann equation

$$\begin{cases} \frac{du}{d\tau} + J_t \frac{du}{dt} = 0, \\ u(\tau, 0) \in L_0,\ u(\tau, 1) \in L_1,\ \int u^*\omega < \infty \end{cases}$$

with the asymptotic condition and the homotopy condition

$$u(-\infty, \cdot) \equiv p, \quad u(\infty, \cdot) \equiv q; \quad [u] = B.$$

We then define $\mathcal{M}(p, q; B) = \widetilde{\mathcal{M}}(p, q; B)/\mathbb{R}$ the quotient by the τ-translations and a collection of *rational* numbers $n(p, q; J, B) = \#(\mathcal{M}(p, q; J, B))$ whenever the expected dimension of $\mathcal{M}(p, q; B)$ is zero. Finally we define the Floer 'boundary' map $\partial\colon\ CF(L_0, L_1; \ell_0) \to CF(L_0, L_1; \ell_0)$ by the sum

$$\partial([p, w]) = \sum_{q\in L_0\cap L_1} \sum_{B\in\pi_2(p,q)} n(p, q; J, B)[q, w \# B]. \tag{4.4}$$

When a Hamiltonian isotopy $\{L'_s\}_{0\le s\le 1}$ is given one also considers the non-autonomous version of the Floer equation

$$\begin{cases} \frac{du}{d\tau} + J_{t,\rho(\tau)} \frac{du}{dt} = 0, \\ u(\tau, 0) \in L,\ u(\tau, 1) \in L'_{\rho(\tau)}, \end{cases}$$

as done in [Oh1] where $\rho\colon \mathbb{R} \to [0,1]$ is a smooth function with $\rho(-\infty) = 0$, $\rho(\infty) = 1$ such that ρ' is compactly supported and define the Floer 'chain' map

$$h\colon\ CF^*(L_0, L_0') \to CF^*(L_1, L_1').$$

However unlike the closed case or the exact case, many things go wrong when one asks for the property $\partial \circ \partial = 0$ or $\partial h + h\partial = 0$ especially over the rational coefficients, and even when $HF^*(L, \phi_H^1(L))$ is defined, it is not isomorphic to the classical cohomology $H^*(L)$.

In the next three subsections, we explain how to overcome these troubles and describe the spectral sequence relating $HF^*(L, \phi_H^1(L))$ to $H^*(L)$ when the former is defined. All the results in these subsections are joint work with H. Ohta and K. Ono that appeared in [FOOO], unless otherwise said. We refer to Ohta's article [Ot] for a more detailed survey on the work from [FOOO].

4.3. Orientation. We first recall the following definition from [FOOO].

Definition 4.4. A submanifold $L \subset M$ is called *relatively spin* if it is orientable and there exists a class $st \in H^2(M, \mathbb{Z}_2)$ such that $st|_L = w_2(TL)$ for the Stiefel–Whitney class $w_2(TL)$ of TL. A pair (L_0, L_1) is relatively spin if there exists a class $st \in H^2(M, \mathbb{Z}_2)$ satisfying $st|_{L_i} = w_2(TL_i)$ for each $i = 0,\ 1$.

We fix such a class $st \in H^2(M, \mathbb{Z}_2)$ and a triangulation of M. Denote by $M^{(k)}$ its k-skeleton. There exists a unique rank 2 vector bundle $V(st)$ on $M^{(3)}$ with $w^1(V(st)) = 0,\ w^2(V(st)) = st$. Now suppose that L is relatively spin and $L^{(2)}$ be the 2-skeleton of L. Then $V \oplus TL$ is trivial on the 2-skeleton of L.

Definition 4.5. We define a (M, st)-relative spin structure of L to be a spin structure of the restriction of the vector bundle $V \oplus TL$ to $L^{(2)}$.

The following theorem was proved independently by de Silva [Si] and in [FOOO].

Theorem 4.6. *The moduli space of pseudo-holomorphic discs is orientable if $L \subset (M, \omega)$ is relatively spin Lagrangian submanifold. Furthermore the choice of relative spin structure on L canonically determines an orientation on the moduli space $\mathcal{M}(L; \beta)$ of holomorphic discs for all $\beta \in \pi_2(M, L)$.*

For the orientations on the Floer moduli spaces the following theorem was proved in [FOOO].

Theorem 4.7. *Let $J = \{J_t\}_{0 \le t \le 1}$ and suppose that a pair of Lagrangian submanifolds (L_0, L_1) are (M, st)-relatively spin. Then for any $p, q \in L_0 \cap L_1$ and $B \in \pi_2(p, q)$ the Floer moduli space $\mathcal{M}(p, q; B)$ is orientable. Furthermore a choice of relative spin structures for the pair (L_0, L_1) determines an orientation on $\mathcal{M}(p, q; B)$.*

One can amplify the orientation to the moduli space of pseudo-holomorphic polygons $\mathcal{M}(\mathcal{L}, \vec{p}; B)$, where $\mathcal{L} = (L_0, L_1, \dots, L_k)$ and $\vec{p} = (p_{01}, p_{12}, \dots, p_{k0})$ with $p_{ij} \in L_i \cap L_j$, and extend the construction to the setting of A_∞-category [Fu1]. We refer to [Fu2] for a more detailed discussion on this.

4.4. Obstruction and A_∞-structure. Let (L_0, L_1) be a relatively spin pair with $L_0 \pitchfork L_1$ and fix a (M, st)-relatively spin structure on each L_i. To convey the appearance of obstruction to the boundary property $\partial\partial = 0$ in a coherent way, we assume in this survey, for simplicity, that all the Floer moduli spaces involved in the construction are transverse and so the expected dimension is the same as the actual dimension. For example, this is the case for monotone Lagrangian submanifolds at least for the Floer moduli spaces of dimension 0, 1 and 2. However, we would like to emphasize that we have to use the machinery of *Kuranishi structure* introduced in [FOn] in the level of *chain* to properly treat the transversality problem for the general case, whose detailed study we refer to [FOOO].

We compute $\partial\partial([p, w])$. According to the definition (4.4) of the map ∂ we have the formula for its matrix coefficients

$$\langle \partial\partial[p, w], [r, w \# B]\rangle = \sum_{q\in L_0\cap L_1} \sum_{B=B_1\#B_2\in\pi_2(p,r)} n(p, q; B_1)n(q, r; B_2) \qquad (4.5)$$

where $B_1 \in \pi_2(p, q)$ and $B_2 \in \pi_2(q, r)$. To prove $\partial\partial = 0$, one needs to prove $\langle \partial\partial[p, w], [r, w \# B]\rangle = 0$ for all pairs $[p, w]$, $[r, w \# B]$. On the other hand it follows from the definition that each summand $n(p, q; B_1)n(q, r; B_2)$ is nothing but the number of broken trajectories lying in $\mathcal{M}(p, q; B_1) \# \mathcal{M}(q, r; B_2)$. The way how Floer [Fl1] proved the vanishing of (4.5) under the assumption that

$$L_0 = L,\ L_1 = \phi_H^1(L);\ \pi_2(M, L_i) = 0 \qquad (4.6)$$

is to construct a suitable compactification of the one-dimensional (smooth) moduli space $\mathcal{M}(p, r; B) = \widetilde{\mathcal{M}}(p, r; B)/\mathbb{R}$ in which the broken trajectories of the form $u_1 \# u_2$ comprise *all* the boundary components of the compactified moduli space. By definition, the expected dimension of $\mathcal{M}(p, r; B)$ is one and so the compactified moduli space becomes a compact one-dimensional manifold. Then $\partial\partial = 0$ follows.

As soon as one goes beyond Floer's case (4.6), one must consider the problems of *a priori energy bound* and *bubbling-off discs*. As in the closed case the Novikov ring is introduced to solve the problem of energy bounds. On the other hand, bubbling-off-discs is a new phenomenon which is that of codimension one and can indeed occur in the boundary of the compactification of Floer moduli spaces.

To handle the problem of bubbling-off discs, Fukaya–Oh–Ohta–Ono [FOOO] associated a structure of filtered A_∞-algebra $(C, \mathfrak{m})$ *with non-zero* $\mathfrak{m}_0$*-term* in general to each compact Lagrangian submanifold. The notion of A_∞-structure was first introduced by Stasheff [St]. We refer to [GJ] for an exposition close to ours with different sign conventions. The above mentioned obstruction is closely related to the non-vanishing of $\mathfrak{m}_0$ in this A_∞-structure. A description of this obstruction is now in order.

Let C be a graded R-module where R is the coefficient ring. In our case R will be $\Lambda_{0,\mathrm{nov}}$. We denote by $C[1]$ its suspension defined by $C[1]^k = C^{k+1}$. We denote by $\deg(x) = |x|$ the degree of $x \in C$ before the shift and by $\deg'(x) = |x|'$ that after

the degree shifting, i.e., $|x|' = |x| - 1$. Define the *bar complex* $B(C[1])$ by

$$B_k(C[1]) = (C[1])^{k\otimes}, \quad B(C[1]) = \bigoplus_{k=0}^{\infty} B_k(C[1]).$$

Here $B_0(C[1]) = R$ by definition. We provide the degree of elements of $B(C[1])$ by the rule

$$|x_1 \otimes \cdots \otimes x_k|' := \sum_{i=1}^{k} |x_i|' = \sum_{i=1}^{k} |x_i| - k \tag{4.7}$$

where $|\cdot|'$ is the shifted degree. The ring $B(C[1])$ has the structure of a *graded coalgebra*.

Definition 4.8. The structure of a (strong) A_∞*-algebra* is a sequence of R-module homomorphisms

$$\mathfrak{m}_k \colon B_k(C[1]) \to C[1], \quad k = 1, 2, \ldots,$$

of degree +1 such that the coderivation $d = \sum_{k=1}^{\infty} \widehat{\mathfrak{m}}_k$ satisfies $dd = 0$, which is called the A_∞*-relation*. Here we denote by $\widehat{\mathfrak{m}}_k \colon B(C[1]) \to B(C[1])$ the unique extension of $\mathfrak{m}_k$ as a coderivation on $B(C[1])$. A *filtered* A_∞*-algebra* is an A_∞-algebra with a filtration for which $\mathfrak{m}_k$ are continuous with respect to the induce non-Archimedean topology.

In particular, we have $\mathfrak{m}_1\mathfrak{m}_1 = 0$ and so it defines a complex $(C, \mathfrak{m}_1)$. We define the $\mathfrak{m}_1$-cohomology by

$$H(C, \mathfrak{m}_1) = \ker \mathfrak{m}_1 / \operatorname{im} \mathfrak{m}_1. \tag{4.8}$$

A *weak* A_∞*-algebra* is defined in the same way, except that it also includes

$$\mathfrak{m}_0 \colon R \to B(C[1]).$$

The first two terms of the A_∞-relation for a weak A_∞-algebra are given as

$$\mathfrak{m}_1(\mathfrak{m}_0(1)) = 0 \tag{4.9}$$

$$\mathfrak{m}_1\mathfrak{m}_1(x) + (-1)^{|x|'}\mathfrak{m}_2(x, \mathfrak{m}_0(1)) + \mathfrak{m}_2(\mathfrak{m}_0(1), x) = 0. \tag{4.10}$$

In particular, for the case of weak A_∞-algebras, $\mathfrak{m}_1$ will not necessarily satisfy the boundary property, i.e., $\mathfrak{m}_1\mathfrak{m}_1 \neq 0$ in general.

The way how a weak A_∞-algebra is attached to a Lagrangian submanifold $L \subset (M, \omega)$ arises as an A_∞-deformation of the classical singular cochain complex including the instanton contributions. In particular, when there is no instanton contribution as in the case $\pi_2(M, L) = 0$, it will reduce to an A_∞-deformation of the singular cohomology in the chain level including all possible higher Massey product. One

outstanding circumstance arises in relation to the *quantization* of rational homotopy theory on the cotangent bundle T^*N of a compact manifold N. In this case the authors proved in [FOh] that the A_∞-subcategory 'generated' by such graphs is literally isomorphic to a certain A_∞-category constructed by the Morse theory of *graph flows*.

We now describe the basic A_∞-operators $\mathfrak{m}_k$ in the context of A_∞-algebra of Lagrangian submanifolds. For a given compatible almost complex structure J consider the moduli space of stable maps of genus zero

$$\begin{aligned}&\mathcal{M}_{k+1}(\beta; L)\\ &\quad = \{(w, (z_0, z_1, \dots, z_k)) \mid \overline{\partial}_J w = 0,\ z_i \in \partial D^2, [w] = \beta \text{ in } \pi_2(M, L)\}/\sim,\end{aligned}$$

where $\sim$ is the conformal reparameterization of the disc D^2. The expected dimension of this space is given by

$$n + \mu(\beta) - 3 + (k+1) = n + \mu(\beta) + k - 2. \tag{4.11}$$

Now given k chains

$$[P_1, f_1], \dots, [P_k, f_k] \in C_*(L)$$

of L considered as *currents* on L, we put the cohomological grading $\deg P_i = n - \dim P_i$ and consider the fiber product

$$ev_0 \colon \mathcal{M}_{k+1}(\beta; L) \times_{(ev_1,\dots,ev_k)} (P_1 \times \cdots \times P_k) \to L.$$

A simple calculation shows that the expected dimension of this chain is given by

$$n + \mu(\beta) - 2 + \sum_{j=1}^{k} (\dim P_j + 1 - n)$$

or equivalently we have the expected degree

$$\deg \left[\mathcal{M}_{k+1}(\beta; L) \times_{(ev_1,\dots,ev_k)} (P_1 \times \cdots \times P_k), ev_0\right] = \sum_{j=1}^{n} (\deg P_j - 1) + 2 - \mu(\beta).$$

For each given $\beta \in \pi_2(M, L)$ and $k = 0, \dots$ we define

$$\mathfrak{m}_{k,\beta}(P_1, \dots, P_k) = \left[\mathcal{M}_{k+1}(\beta; L) \times_{(ev_1,\dots,ev_k)} (P_1 \times \cdots \times P_k), ev_0\right]$$

and $\mathfrak{m}_k = \sum_{\beta \in \pi_2(M,L)} \mathfrak{m}_{k,\beta} \cdot q^\beta$ where $q^\beta = T^{\omega(\beta)} e^{\mu(\beta)/2}$ with T, e formal parameters encoding the area and the Maslov index of β. We provide T with degree 0 and e with 2. Now we denote by $C[1]$ the completion of a *suitably chosen countably generated* cochain complex with $\Lambda_{0,\mathrm{nov}}$ as its coefficients with respect to the non-Archimedean topology. Then it follows that the map $\mathfrak{m}_k \colon C[1]^{\otimes k} \to C[1]$ is well defined, has degree 1 and is continuous with respect to the non-Archimedean topology. We extend $\mathfrak{m}_k$ as a coderivation $\widehat{\mathfrak{m}}_k \colon BC[1] \to BC[1]$ where $BC[1]$ is the

completion of the direct sum $\bigoplus_{k=0}^{\infty} B^k C[1]$ and where $B^k C[1]$ itself is the completion of $C[1]^{\otimes k}$. $BC[1]$ has a natural filtration defined similarly as 4.3. Finally we take the sum

$$\hat{d} = \sum_{k=0}^{\infty} \widehat{m}_k : \ BC[1] \to BC[1].$$

A main theorem then is the following coboundary property.

Theorem 4.9. *Let L be an arbitrary compact relatively spin Lagrangian submanifold of an arbitrary tame symplectic manifold (M, ω). The coderivation $\hat{d}$ is a continuous map that satisfies the A_∞-relation $\hat{d}\hat{d} = 0$, and so $(C, \mathfrak{m})$ is a filtered weak A_∞-algebra.*

One might want to consider the homology of this huge complex, but if one naively takes the homology of this complex itself, it will end up with getting a trivial group, which is isomorphic to the ground ring $\Lambda_{0,\mathrm{nov}}$. This is because the A_∞-algebra associated to L in [FOOO] has the (*homotopy*) *unit*: if an A_∞-algebra has a unit, the homology of $\hat{d}$ is isomorphic to its ground ring.

A more geometrically useful homology relevant to the Floer homology is the $\mathfrak{m}_1$-homology (4.8) in this context, which is the Bott–Morse version of the Floer cohomology for the pair (L, L). However in the presence of $\mathfrak{m}_0$, $\mathfrak{m}_1\mathfrak{m}_1 = 0$ no longer holds in general. Motivated by Kontsevich's suggestion [K2], this led Fukaya–Oh–Ohta–Ono to consider deforming Floer's original definition by a bounding chain of the obstruction cycle arising from bubbling-off discs. One can always deform the given (filtered) A_∞-algebra $(C, \mathfrak{m})$ by an element $b \in C[1]^0$ by re-defining the A_∞-operators as

$$\mathfrak{m}_k^b(x_1, \ldots, x_k) = \mathfrak{m}(e^b, x_1, e^b, x_2, e^b, x_3, \ldots, x_k, e^b)$$

and taking the sum $\hat{d}^b = \sum_{k=0}^{\infty} \widehat{\mathfrak{m}}_k^b$. This defines a new weak A_∞-algebra in general. Here we simplify the notation by writing

$$e^b = 1 + b + b \otimes b + \cdots + b \otimes \cdots \otimes b + \cdots .$$

Note that each summand in this infinite sum has degree 0 in $C[1]$ and converges in the non-Archimedean topology if b has positive valuation, i.e., $v(b) > 0$.

Proposition 4.10. *For the A_∞-algebra $(C, \mathfrak{m}_k^b)$, $\mathfrak{m}_0^b = 0$ if and only if b satisfies*

$$\sum_{k=0}^{\infty} \mathfrak{m}_k(b, \ldots, b) = 0. \tag{4.12}$$

This equation is a version of the Maurer–Cartan equation *for the filtered A_∞-algebra.*

Definition 4.11. Let $(C, \mathfrak{m})$ be a filtered weak A_∞-algebra in general and $BC[1]$ be its bar complex. An element $b \in C[1]^0 = C^1$ is called a *bounding cochain* if it satisfies the equation (4.12) and $v(b) > 0$.

In general a given A_∞-algebra may or may not have a solution to (4.12).

Definition 4.12. A filtered weak A_∞-algebra is called *unobstructed* if the equation (4.12) has a solution $b \in C[1]^0 = C^1$ with $v(b) > 0$.

One can define a notion of homotopy equivalence between two bounding cochains as described in [FOOO]. We denote by $\mathcal{M}(L)$ the set of equivalence classes of bounding cochains of L.

Once the A_∞-algebra is attached to each Lagrangian submanifold L, we then construct an A_∞-*bimodule* $C(L, L')$ for the pair by considering operators

$$\mathfrak{n}_{k_1,k_2} \colon C(L, L') \to C(L, L')$$

defined similarly to $\mathfrak{m}_k$: A typical generator of $C(L, L')$ has the form

$$P_{1,1} \otimes \cdots \otimes P_{1,k_1} \otimes [p, w] \otimes P_{2,1} \otimes \cdots \otimes P_{2,k_2}$$

and then the image $\mathfrak{n}_{k_1,k_2}$ thereof is given by

$$\sum_{[q,w']} \left[\left(\mathcal{M}([p, w], [q, w']; P_{1,1}, \ldots, P_{1,k_1}; P_{2,1}, \ldots, P_{2,k_2}), ev_\infty \right) \right] [q, w'].$$

Here $\mathcal{M}([p, w], [q, w']; P_{1,1}, \ldots, P_{1,k_1}; P_{2,1}, \ldots, P_{2,k_2})$ is the Floer moduli space

$$\mathcal{M}([p, w], [q, w']) = \bigcup_{[q,w']=[q,w\#B]} \mathcal{M}(p, q; B)$$

cut-down by intersecting with the given chains $P_{1,i} \subset L$ and $P_{2,j} \subset L'$, and the evaluation map

$$ev_\infty \colon \mathcal{M}([p, w], [q, w']; P_{1,1}, \ldots, P_{1,k_1}; P_{2,1}, \ldots, P_{2,k_2}) \to \mathrm{Crit}\mathcal{A}$$

is defined by $ev_\infty(u) = u(+\infty)$.

Theorem 4.13. *Let (L, L') be an arbitrary relatively spin pair of compact Lagrangian submanifolds. Then the family $\{\mathfrak{n}_{k_1,k_2}\}$ defines a left $(C(L), \mathfrak{m})$ and right $(C(L'), \mathfrak{m}')$ filtered A_∞-bimodule structure on $C(L, L')$.*

In other words, each of the map $\mathfrak{n}_{k_1,k_2}$ extends to a A_∞-bimodule homomorphism $\hat{\mathfrak{n}}_{k_1,k_2}$ and if we take the sum

$$\hat{d} := \sum_{k_1,k_2} \hat{\mathfrak{n}}_{k_1,k_2} \colon C(L, L') \to C(L, L'),$$

$\hat{d}$ satisfies the following coboundary property.

Proposition 4.14. *The map $\hat{d}$ is a continuous map and satisfies $\hat{d}\hat{d} = 0$.*

Again this complex is too big for computational purposes and we would like to consider the Floer homology by restricting the A_∞-bimodule to a much smaller complex, an ordinary Λ_{nov}-module $CF(L, L')$. However Floer's original definition again meets obstruction coming from the obstruction cycles of either L_0, L_1 or of both. We need to deform Floer's 'boundary' map δ using suitable bounding cochains of L, L'. The bimodule $C(L, L')$ is introduced to perform this deformation coherently.

In the case where both L, L' are unobstructed, we can carry out this deformation of $\mathfrak{n}$ by bounding chains $b_1 \in \mathcal{M}(L)$ and $b_2 \in \mathcal{M}(L')$ similarly as $\mathfrak{m}^b$ above. Symbolically we can write the new operator as

$$\delta^{b_1,b_2}(x) = \mathfrak{n}(e^{b_1}, x, e^{b_2}).$$

Theorem 4.15. *For each $b_1 \in \mathcal{M}(L)$ and $b_2 \in \mathcal{M}(L')$, the map δ^{b_1,b_2} defines a continuous map $\delta^{b_1,b_2}\colon CF(L, L') \to CF(L, L')$ that satisfies $\delta^{b_1,b_2}\delta^{b_1,b_2} = 0$.*

This theorem enables us to define the *deformed Floer cohomology*.

Definition 4.16. For each $b \in \mathcal{M}(L)$ and $b' \in \mathcal{M}(L')$, we define the (b, b')-*Floer cohomology* of the pair (L, L') by

$$HF((L, b), (L', b'); \Lambda_{\text{nov}}) = \frac{\ker \delta^{b_1,b_2}}{\text{im } \delta^{b_1,b_2}}.$$

Theorem 4.17. *The above cohomology remains isomorphic under the Hamiltonian isotopy of L, L' and under the homotopy of bounding cochains b, b'.*

We refer to [FOOO] and its revised version for all the details of algebraic language needed to make the statements in the above theorems precise.

4.5. Spectral sequence. The idea of spectral sequence is quite simple to describe. One can follow more or less the standard construction of the spectral sequence on the filtered complex, as e.g. in [Mc]. One trouble to overcome in the construction of the spectral sequence on $(C(L), \delta)$ or $(C(L, L'), \delta)$ is that the general Novikov ring, in particular $\Lambda_{0,\text{nov}}$ is *not* Noetherian and so the standard theorems on Noetherian modules cannot be applied. In addition, the Floer complex is not bounded above which also makes the proof of convergence of the spectral sequence somewhat tricky. We refer to [FOOO] for a complete discussion on the construction of the spectral sequence and the study of their convergences.

However for the case of monotone Lagrangian submanifolds, the Novikov ring becomes a field and the corresponding spectral sequence is much more simplified as originally carried out by Oh [Oh3] by a crude analysis of thick-thin decomposition of Floer moduli spaces as two Lagrangian submanifolds collapse to one. Then the geometric origin of the spectral sequence is the decomposition of the Floer boundary map δ into $\delta = \delta_0 + \delta_1 + \delta_2 + \cdots$ where each δ_i is the contribution coming from the Floer trajectories of a given symplectic area in a way that the corresponding area is

increasing as $i \to \infty$. Here δ_0 is the contribution from the classical cohomology. In general this sequence may not stop at a finite stage but it does for monotone Lagrangian submanifolds. In this regard, we can roughly state the following general theorem:

There exists a spectral sequence whose E^2-term is isomorphic to the singular cohomology $H^*(L)$ and which converges to the Floer cohomology $HF^*(L, L)$.

See [Oh3] and [FOOO] for the details of the monotone case and of the general case respectively. The above decomposition also provides an algorithm to utilize the spectral sequence in examples, especially when the Floer cohomology is known as for the case of Lagrangian submanifolds in $\mathbb{C}^n$. Here are some sample results.

Theorem 4.18 (Theorem II [Oh3]). *Let (M, ω) be a tame symplectic manifold with* $\dim M \geq 4$. *Let L be a compact monotone Lagrangian submanifold of M and ϕ be a Hamiltonian diffeomorphism of (M, ω) such that $L \pitchfork \phi(L)$. Then the following assertions hold:*

1. *If $\Sigma_L \geq n + 2$, $HF^k(L, \phi(L); \mathbb{Z}_2) \cong H^k(L; \mathbb{Z}_2)$ for all $k \mod \Sigma_L$.*
2. *If $\Sigma_L = n + 1$, the same is true for $k \neq 0, n \mod n + 1$.*

Theorem 4.19 (Theorem III [Oh3]).[1] *Let $L \subset \mathbb{C}^n$ be a compact monotone Lagrangian torus. Then we have $\Sigma_L = 2$ provided $1 \leq n \leq 24$.*

A similar consideration, using a more precise form of the spectral sequence from [FOOO], proves

Theorem 4.20. *Suppose that $L \subset \mathbb{C}^n$ is a compact Lagrangian embedding with $H^2(L; \mathbb{Z}_2) = 0$. Then its Maslov class μ_L is not zero.*

The following theorem can be derived from Theorem E [FOOO] which should be useful for the study of intersection properties of special Lagrangian submanifolds on Calabi–Yau manifolds.

Theorem 4.21. *Let M be a Calabi–Yau manifold and L be an unobstructed Lagrangian submanifold with its Maslov class $\mu_L = 0$ in $H^1(L; \mathbb{Z})$. Then we have $HF^i(L; \Lambda_{0,\mathrm{nov}}) \neq 0$ for $i = 0, \dim L$.*

For example, any special Lagrangian homology sphere satisfies all the hypotheses required in this theorem. Using this result combined with some Morse theory argument, Thomas and Yau [TY] proved the following uniqueness result of special Lagrangian homology sphere in its Hamiltonian isotopy class

Theorem 4.22 (Thomas-Yau). *For any Hamiltonian isotopy class of embedded Lagrangian submanifold L with $H^*(L) \cong H^*(S^n)$ there exists at most one smooth special Lagrangian representative.*

Biran [Bi] also used this spectral sequence for the study of geometry of Lagrangian skeletons and polarizations of Kähler manifolds.

[1] *Added in proof.* In the revised version of [FOOO] the dimensional restriction $1 \leq n \leq 24$ was removed following the scheme suggested by Biran [Bi].

5. Displaceable Lagrangian submanifolds

Definition 5.1. We call a compact Lagrangian submanifold $L \subset (M, \omega)$ displaceable if there exists a Hamiltonian isotopy ϕ_H such that $L \cap \phi_H^1(L) = \emptyset$.

One motivating question for studying such Lagrangian submanifolds is the following well-known folklore conjecture in symplectic geometry.

Conjecture 5.2 (Maslov Class Conjecture). Any compact Lagrangian embedding in $\mathbb{C}^n$ has non-zero Maslov class.

Polterovich [P] proved the conjecture in dimension $n = 2$ whose proof uses a loop γ realized by the boundary of Gromov's holomorphic disc constructed in [Gr]. Viterbo proved this conjecture for any Lagrangian torus in $\mathbb{C}^n$ by a different method using the critical point theory on the free loop spaces of $\mathbb{C}^n$ [V1]. Also see Theorem 4.20 in the previous section for L with $H^2(L; \mathbb{Z}_2) \neq 0$.

It follows from the definition that $HF^*(L, \phi_H^1(L)) = 0$ for a displaceable Lagrangian submanifold L *whenever* $HF^*(L, \phi_H^1(L))$ *is defined*. An obvious class of displaceable Lagrangian submanifolds are those in $\mathbb{C}^n$. This simple observation, when combined with the spectral sequence described in the previous section, provides many interesting consequences on the symplectic topology of such Lagrangian submanifolds as illustrated by Theorem 4.18 and 4.19.

Some further amplification of this line of reasoning was made by Biran and Cieliebak [BC] for the study of topology of Lagrangian submanifolds in (complete) *sub-critical Stein manifolds* (V, J) or a symplectic manifold M with such V as a factor. They cooked up some class of Lagrangian submanifolds in such symplectic manifolds with suitable condition on the first Chern class of M under which the Lagrangian submanifolds become monotone and satisfy the hypotheses in Theorem 4.18. Then applying this theorem, they derived restrictions on the topology of such Lagrangian submanifolds, e.g., some *cohomological sphericality* of such Lagrangian submanifolds (see Theorem 1.1 [BC]).

Recently Fukaya [Fu4] gave a new construction of the A_∞-structure described in the previous section as a deformation of the differential graded algebra of the de Rham complex of L associated to a natural solution to the Maurer–Cartan equation of the Batalin–Vilkovisky structure discovered by Chas and Sullivan [CS] on the loop space. In this way, Fukaya combined Gromov's and Polterovich's pseudo-holomorphic curve approach and Viterbo's loop space approach [V1] and proved several new results on the structure of Lagrangian embeddings in $\mathbb{C}^n$. The following are some sample results proven by this method [Fu4]:

1. If L is spin and aspherical in $\mathbb{C}^n$ then a finite cover $\widetilde{L}$ of L is homotopy equivalent to a product $S^1 \times \widetilde{L}'$. Moreover the Maslov index of $[S^1] \times [\text{point}]$ is 2.

2. If $S^1 \times S^{2n}$ is embedded as a Lagrangian submanifold of $\mathbb{C}^{2n+1}$, then the Maslov index of $[S^1] \times [\text{point}]$ is 2.

There is also the symplectic field theory approach to the proof of the first result above for the case of torus $L = T^n$ as Eliashberg explained to the authors [El2]. Eliashberg's scheme has been further detailed by Cieliebak and Mohnke [CM]. The first result for T^n is an affirmative answer to Audin's question [Au] on the minimal Maslov number of the embedded Lagrangian torus in $\mathbb{C}^n$ for general n. Previously this was known only for $n = 2$, [P], [V1], and for monotone Lagrangian tori [Oh3] (see Theorem 4.19).

6. Applications to mirror symmetry

Mirror symmetry discovered in super-string theory attracted much attention from many (algebraic) geometers since it made a remarkable prediction on the relation between the number of rational curves on a Calabi–Yau 3-fold M and the deformation theory of complex structures of another Calabi–Yau manifold $M^\dagger$.

6.1. Homological mirror symmetry. Based on Fukaya's construction of the A_∞-category of symplectic manifolds [Fu1], Kontsevich [K1] proposed a conjecture on the relation between the category $\mathrm{Fuk}(M)$ of (M, ω) and the derived category of coherent sheaves $\mathrm{Coh}(M^\dagger)$ of $M^\dagger$, and extended the mirror conjecture in a more conceptual way. This extended version is called the *homological mirror symmetry*, which is closely related to the D-brane duality studied much in physics. Due to the obstruction phenomenon we described in §3.4, the original construction in [Fu1] requires some clarification of the definition of $\mathrm{Fuk}(M)$. The necessary modification has been completed in [FOOO], [Fu2].

For the rest of this subsection, we will formulate a precise mathematical conjecture of homological mirror symmetry. Let (M, ω) be an integral symplectic manifold, i.e., one with $[\omega] \in H^2(M; \mathbb{Z})$. For such (M, ω), we consider a family of complexified symplectic structures $M_\tau = (M, -\sqrt{-1}\tau\omega)$ parameterized by $\tau \in \mathfrak{h}$ where $\mathfrak{h}$ is the upper half plane. The mirror of this family is expected to be a family of complex manifolds $M_q^\dagger$ parameterized by $q = e^{\sqrt{-1}\tau} \in D^2 \setminus \{0\}$, the punctured disc. Suitably 'formalizing' this family at 0, we obtain a scheme $\mathfrak{M}^\dagger$ defined over the ring $\mathbb{Q}[[q]][q^{-1}]$. We identify $\mathbb{Q}[[q]][q^{-1}]$ with a sub-ring of the universal Novikov ring Λ_{nov} defined in Subsection 4.1. The ext group $\mathrm{Ext}(\mathcal{E}_0, \mathcal{E}_1)$ between the coherent sheaves $\mathcal{E}_i$ on $\mathfrak{M}^\dagger$ is a module over $\mathbb{Q}[[q]][q^{-1}]$.

We consider the quadruple $\mathcal{L} = (L, s, d, [b])$, which we call a *Lagrangian brane*, that satisfies the following data:

1. L a Lagrangian submanifold of M such that the Maslov index of L is zero and $[\omega] \in H^2(M, L; \mathbb{Z})$. We also enhance L with flat complex line bundle on it.
2. s is a spin structure of L.
3. d is a grading in the sense of [K1], [Se1].
4. $[b] \in \mathcal{M}(L)$ is a bounding cochain described in Subsection 4.4.

Conjecture 6.1. To each Lagrangian brane $\mathcal{L}$ as above, we can associate an object $\mathcal{E}(\mathcal{L})$ of the derived category of coherent sheaves on the scheme $\mathfrak{M}^\dagger$ so that the following holds:

1. There exists a canonical isomorphism

$$HF(\mathcal{L}_1, \mathcal{L}_2) \cong \mathrm{Ext}(\mathcal{E}(\mathcal{L}_1), \mathcal{E}(\mathcal{L}_2)) \otimes_{\mathbb{Q}[[q]][q^{-1}]} \Lambda_{\mathrm{nov}}.$$

2. The isomorphism in 1. is functorial: namely the product of Floer cohomology is mapped to the Yoneda product of the Ext group by the isomorphism in 1.

The correct Floer cohomology $HF(\mathcal{L}_1, \mathcal{L}_2)$ used in this formulation of the conjecture is given in [FOOO] (see §3.4 for a brief description). The spin structure in $\mathcal{L}$ is needed to define orientations on the various moduli spaces involved in the definition of Floer cohomology, and the grading d is used to define an absolute integer grading on $HF(\mathcal{L}_1, \mathcal{L}_2)$. We refer the reader to [FOOO] §1.4, [Fu3] for the details of construction and for more references.

We now provide some evidences for this conjecture. A conjecture of this kind was first made by Kontsevich in [K1] for the case of an elliptic curve M, which is further explored by Polischchuk–Zaslow [PZ], and by Fukaya in [Fu3] for the case when M is a torus (and so $M^\dagger$ is also a torus) and $L \subset M$ is an affine sub-torus. In fact, in these cases one can use the convergent power series for the formal power series or the Novikov ring. Kontsevich–Soibelman [KS] gave an alternative proof, based on the adiabatic degeneration result of the authors [FOh], for the case where L is an etalé covering of the base torus of the Lagrangian torus fibration $M = T^{2n} \to T^n$. Seidel proved Conjecture 5.1 for the quartic surface M [Se2].

6.2. Toric Fano and Landau–Ginzburg correspondence. So far we have discussed the case of Calabi–Yau manifolds (or a symplectic manifold (M, ω) with $c_1(M) = 0$). The other important case that physicists studied much is the case of toric Fano manifolds, which physicists call the correspondence between the σ-model and the Landau–Ginzburg model. Referring readers to [Ho] and [HV] for detailed physical description of this correspondence, we briefly describe an application of the machinery developed in [FOOO] for an explicit calculation of Floer cohomology of Lagrangian torus orbits of toric Fano manifolds. We will focus on the correspondence of the A-model of a toric Fano manifold and the B-model of Landau–Ginzburg model of its mirror. We refer to [HIV] for the other side of the correspondence between the toric Fano B-model and the Landau–Ginzburg A-model.

According to [FOOO] the obstruction cycles of the filtered A_∞-algebra associated to a Lagrangian submanifold is closely related to $\mathfrak{m}_0$. This $\mathfrak{m}_0$ is defined by a collection of the (co)chains $[\mathcal{M}_1(\beta), ev_0]$ for all $\beta \in \pi_2(M, L)$. More precisely, we have

$$\mathfrak{m}_0(1) = \sum_{\beta \in \pi_2(M,L)} [\mathcal{M}_1(\beta), ev_0] \cdot T^{\omega(\beta)} q^{\mu(\beta)/2} \in C^*(L) \otimes \Lambda_{0,\mathrm{nov}}. \tag{6.1}$$

This is the sum of all genus zero instanton contributions with one marked point.

On the other hand, based on a B-model calculations, Hori [Ho], Hori–Vafa [HV] proposed some correspondence between the instanton contributions of the A-model of toric Fano manifolds and the Landau–Ginzburg potential of the B-model of its mirror. This correspondence was made precise by Cho and Oh [CO]. A description of this correspondence is now in order. First they proved the following

Theorem 6.2. $[\mathcal{M}_1(\beta), ev_0] = [L]$ *as a chain, for every* $\beta \in \pi_2(M, L)$ *with* $\mu(\beta) = 2$ *and so* $\mathfrak{m}_0(1) = \lambda[L]$ *for some* $\lambda \in \Lambda_{0,\mathrm{nov}}$.

It had been previously observed in Addenda of [Oh1] for the monotone case that Floer cohomology $HF^*(L, L)$ is defined even when the minimal Maslov number $\Sigma_L = 2$. Using the same argument Cho and Oh proved that $HF^*(L, L; \Lambda_{0,\mathrm{nov}})$ is well defined for the torus fibers of toric Fano manifolds *without deforming* Floer's 'boundary' map, at least for the *convex* case. We believe this convexity condition can be removed. More specifically, they proved $\mathfrak{m}_1\mathfrak{m}_1 = 0$. This is because in (4.10) the last two terms cancel each other *if* $\mathfrak{m}_0(1) = \lambda e$ *is a multiple of the unit* $e = [L]$ and then (4.9) implies that $\mathfrak{m}_0(1)$ is a $\mathfrak{m}_1$-cycle. We refer the reader to the revision of [FOOO] for a further discussion on this case, in which any filtered A_∞-algebra deformable to such a case is called *weakly unobstructed*.

In fact, Cho and Oh obtained the explicit formula

$$\mathfrak{m}_0(1) = \sum_{i=1}^{N} h^{v_j} e^{\sqrt{-1}\langle v, v_j\rangle} T^{\omega(\beta_j)}[L] \cdot q \tag{6.2}$$

after including flat line bundles attached to L and computing precise formulae for the area $\omega(\beta_j)$'s. Here $h^{v_j} = e^{\sqrt{-1}\langle v, v_j\rangle}$ is the holonomy of the flat line bundle and $\omega(\beta_j)$ was calculated explicitly in [CO]. Denote by $v = (v_1, \ldots, v_N)$ the holonomy vector of the line bundle appearing in the description of linear σ-model [HV].

On the other hand, the Landau–Ginzburg potential is given by the formula

$$\sum_{i=1}^{N} \exp(-y_i - \langle \Theta, v_i\rangle) =: W(\Theta)$$

for the mirror of the given toric manifold (see [HV] for example).

Theorem 6.3 ([CO]). *Let* $A \in \mathfrak{t}^*$ *and denote* $\Theta = A - \sqrt{-1}v$. *We denote by* $\mathfrak{m}^\Theta$ *the* A_∞*-operators associated to the torus fiber* $T_A = \pi^{-1}(A)$ *coupled with the flat line bundle whose holonomy vector is given by* $v \in (S^1)^N$ *over the toric fibration* $\pi : X \to \mathfrak{t}^*$. *Under the substitution of* $T^{2\pi} = e^{-1}$ *and ignoring the harmless grading parameter* q, *we have the exact correspondences*

$$\mathfrak{m}_0^\Theta \longleftrightarrow W(\Theta), \tag{6.3}$$

$$\mathfrak{m}_1^\Theta(pt) \longleftrightarrow dW(\Theta) = \sum_{j=1}^{n} \frac{\partial W}{\partial \Theta_j}(\Theta) d\Theta_j \tag{6.4}$$

under the mirror map given in [HV].

Combined with a theorem from [CO] which states that $HF^*(\mathcal{L}, \mathcal{L}) \cong H^*(L; \mathbb{C}) \otimes \Lambda_{0,\mathrm{nov}}$ whenever $\mathfrak{m}_1^{\Theta}(pt) = 0$, this theorem confirms the prediction made by Hori [Ho], Hori–Vafa [HV] about the Floer cohomology of Lagrangian torus fibers. This theorem has been further enhanced by Cho [Cho] who relates the higher derivatives of W with the higher Massey products $\mathfrak{m}_k^{\Theta}$. For example, Cho proved that the natural product structure on $HF^*(L, L)$ is not isomorphic to the cohomology ring $H^*(T^n) \otimes \Lambda_{0,\mathrm{nov}}$ but isomorphic to the Clifford algebra associated to the quadratic form given by the Hessian of the potential W under the mirror map. This was also predicted by physicists (see [Ho]).

References

[Au] Audin, M., Fibres normaux d'immersions en dimension double, points doubles d'immersions Lagrangiennes et plongements totalement réeles. *Comment. Math. Helv.* **63** (1988), 593–623.

[Ba] Banyaga, A., Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique. *Comment. Math. Helv.* **53** (1978), 174–227.

[Bi] Biran, P., Lagrangian non-intersections. Preprint, 2004; arXiv:math.SG/0412110.

[BC] Biran, P., Cieliebak, K., Lagrangian embeddings into subcritical Stein manifolds. *Israel J. Math.* **127** (2002), 221–244.

[CS] Chas, M., Sullivan D., String topology. Preprint, 1999, math.GT/9911159.

[Cho] Cho, C.-H., Products of Floer cohomology of Lagrangian torus fibers in toric Fano manifolds. *Comm. Math. Phys.* **260** (2005), 613–640.

[CO] Cho, C.-H., Oh, Y.-G., Floer cohomology and disc instantons of Lagrangian torus fibers in toric Fano manifolds. *Asian J. Math.*, to appear; arXiv:math.SG/0308225.

[CM] Cieliebak, K., Mohnke, K., A talk by Cieliebak in Banach Center, Warsaw 2004.

[EH] Ekeland, I., Hofer, H., Symplectic topology and Hamiltonian dynamics I & II. *Math. Z.* **200** (1990), 355–378; **203** (1990), 553–567.

[El1] Eliashberg, Y., A theorem on the structure of wave fronts and applications in symplectic topology. *Functional Anal. Appl.* **21** (1987), 227–232.

[El2] Eliashberg, Y., private communication, 2001.

[EnP] Entov, M., Polterovich, L., Calabi quasimorphism and quantum homology. *Internat. Math. Res. Notices* **30** (2003), 1635–1676.

[Fa] Fathi, A., Structure of the group of homeomorphisms preserving a good measure on a compact manifold. *Ann. Sci. École Norm. Sup.* **13** (1980), 45–93.

[Fl1] Floer, A., Morse theory for Lagrangian interesections. *J. Differential Geom.* **28** (1988), 513–547.

[Fl2] Floer, A., Symplectic fixed points and holomorphic spheres. *Comm. Math. Phys.* **120** (1989), 575–611.

[Fu1] Fukaya, K., Morse homotopy, A^∞-category, and Floer homologies. In *Proceedings of GARC Workshop on Geometry and Topology* '93, Lecture Notes Ser. 18, Seoul National University, Seoul 1993, 1–102.

[Fu2] Fukaya, K., Floer homology and mirror symmetry. II. In *Minimal surfaces, geometric analysis and symplectic geometry*, Adv. Stud. Pure Math. 34, Math. Soc. Japan, Tokyo 2002, 31–127.

[Fu3] Fukaya, K., Mirror symmetry of abelian varieties and multi-theta functions. *J. Algebraic Geom.* **11** (2002), 393–512.

[Fu4] Fukaya, K., Application of Floer Homology of Langrangian Submanifolds to Symplectic Topology. In *Morse Theoretic Methods in Nonlinear Analysis and in Symplectic Topology* (ed. by P. Biran, O. Cornea, F. Lalonde), NATO Sci. Ser. II 217, Springer-Verlag, Berlin, New York 2006, 231–276.

[FOh] Fukaya, K., Oh, Y.-G., Zero-loop open strings in the cotangent bundle and Morse homotopy. *Asian J. Math.* **1** (1997), 96–180.

[FOOO] Fukaya, K., Oh. Y.-G., Ohta, H., Ono, K., *Lagrangian intersection Floer homology – anomaly and obstruction*. Preprint, 2000, revision, 2006; http://www.math.kyoto-u.ac.jp/~fukaya/.

[FOn] Fukaya, K., Ono, K., Arnold conjecture and Gromov-Witten invariants. *Topology* **38** (1999), 933–1048.

[GJ] Getzler, E., Jones, D. S., A_∞-algebra and cyclic bar complex. *Illinois J. Math.* **34** (1990), 256–283.

[Gr] Gromov, M., Pseudo-holomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), 307–347.

[H] Hofer, H., On the topological properties of symplectic maps. *Proc. Royal Soc. Edinburgh* **115** (1990), 25–38.

[HS] Hofer, H., Salamon, D., Floer homology and Novikov rings. In *The Floer Memorial Volume* (ed. by H. Hofer, C. Taubes, A. Weinstein and E. Zehnder), Birkhäuser, Basel 1995, 483–524.

[Ho] Hori, K., Linear models in supersymmetric D-branes. In *Symplectic Geometry and Mirror Symmetry, Seoul* 2000 (ed. by K. Fukaya, Y.-G. Oh, K. Ono and G. Tian), World Sci. Publishing, River Edge, NJ, 2001, 111–186.

[HIV] Hori, K., Iqbal, A., Vafa, C., D-branes and mirror symmetry. Preprint; hepth/0005247.

[HV] Hori, K., Vafa, C., Mirror symmetry. Preprint, 2000; hep-th/0002222.

[K1] Kontsevich, M., Homological algebra of mirror symmetry. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 120–139.

[K2] Kontsevich, M., Private communication, 1997.

[KS] Kontsevich, M., Soibelman. Y., Homological mirror symmetry and torus fibrations. In *Symplectic Geometry and Mirror Symmetry, Seoul* 2000 (ed. by K. Fukaya, Y.-G. Oh, K. Ono and G. Tian), World Sci. Publishing, River Edge, NJ, 2001, 203–263.

[LT1] Liu, G., Tian, G., Floer homology and Arnold conjecture. *J. Differential Geom.* **49** (1998), 1–74.

[Mc] McCleary, J., *User's Guide to Spectral Sequences*. Math. Lecture Series 12, Publish or Perish, Wilmington, De, 1985.

[N] Novikov, S. P., Multivalued functions and functionals. An analogue of the Morse theory. *Dokl. Akad. Nauk SSSR* **260** (1) (1981), 31–35 (in Russian).

[Oh1] Oh, Y.-G., Floer cohomology of Lagrangian intersections and pseudo-holomorphic discs, I & II. *Comm. Pure and Applied Math.* **46** (1993), 949–994, 995–1012; Addendum ibid. **48** (1995), 1299–1302.

[Oh2] Oh, Y.-G., Relative Floer and quantum cohomology and the symplectic topology of Lagrangian submanifolds. In *Contact and Symplectic Geometry* (ed. by C. B. Thomas), Publ. Newton Inst. 8, Cambridge University Press, Cambridge 1996, 201–267.

[Oh3] Oh, Y.-G., Floer cohomology, spectral sequences, and the Maslov class of Lagrangian embeddings. *Internat. Math. Res. Notices* **7** (1996), 305–346.

[Oh4] Oh, Y.-G., Symplectic topology as the geometry of action functional, I. *J. Differential Geom.* **46** (1997), 499–577.

[Oh5] Oh, Y.-G., Symplectic topology as the geometry of action functional, II. *Comm. Anal. Geom.* **7** (1999), 1–55.

[Oh6] Oh, Y.-G., Chain level Floer theory and Hofer's geometry of the Hamiltonian diffeomorphism group. *Asian J. Math.* **6** (2002), 579–624; Erratum ibid. **7** (2003), 447–448.

[Oh7] Oh, Y.-G., Spectral invariants and length minimizing property of Hamiltonian paths. *Asian J. Math.* **9** (2005), 1–18.

[Oh8] Oh, Y.-G., Construction of spectral invariants of Hamiltonian paths on closed symplectic manifolds. In *The Breadth of Symplectic and Poisson Geometry*, Prog. Math. 232, Birkhäuser, Boston 2005, 525–570.

[Oh9] Oh, Y.-G., Spectral invariants, analysis of the Floer moduli space, and geometry of the Hamiltonian diffeomorphism group. *Duke Math. J.* **130** (2005), 199–295.

[Oh10] Oh, Y.-G., Floer mini-max theory, the Cerf diagram, and the spectral invariants. Preprint, 2004; math.SG/0406449.

[Oh11] Oh, Y.-G., C^0-coerciveness of Moser's problem and smoothing of area preserving diffeomorphisms. Submitted, 2005.

[Oh12] Oh, Y.-G., The group of Hamiltonian homeomorphisms and topological Hamiltonian flows. Submitted, December 2005.

[OM] Oh, Y.-G., Müller, S., The group of Hamiltonian homeomorphisms and C^0-symplectic topology. Submitted, revision December 2005; math.SG/0402210.

[Ot] Ohta, H., Obstruction to and deformation of Lagrangian intersection Floer cohomology. In *Mirror Symmetry and Symplectic Geometry, Seoul* 2000 (ed. by K. Fukaya, Y.-G. Oh, K. Ono and G. Tian), World Sci. Publishing, River Edge, NJ, 2001, 281—309.

[On] Ono, K., On the Arnold conjecture for weakly monotone symplectic manifolds. *Invent. Math.* **119** (1995), 519–537.

[P] Polterovich, L., The Maslov class of Lagrange surfaces and Gromov's pseudo-holomorphic curves. *Trans. Amer. Math. Soc.* **325** (1991), 241–248.

[PZ] Polishchuk, A., Zaslow. E., Categorical mirror symmetry: the elliptic curve. *Adv. Theor. Math. Phys.* **2** (2) (1998), 443–470.

[Ru] Ruan, Y., Virtual neighborhood and pseudo-holomorphic curves. *Turkish J. Math.* **23** (1999), 161–231.

[Sc] Schwarz, M., On the action spectrum for closed symplectically aspherical manifolds. *Pacific J. Math.* **193** (2000), 419–461.

[Se1] Seidel, P., Graded Lagrangian submanifolds. *Bull. Soc. Math. France* **128** (2000), 103–149.

[Se2] Seidel, P., Homological mirror symmetry for the quartic surface. Preprint, 2003; math.SG/0310414.

[Si] de Silva, V., Products in the symplectic Floer homology of Lagrangian intersections. Ph.D. thesis, Oxford University, 1998.

[St] Stasheff, J., Homotopy associativity of H-Spaces I & II. *Trans. Amer. Math. Soc.* **108** (1963), 275–312, 293–312.

[TY] Thomas, R., Yau, S.-T., Special Lagrangians, stable bundles and mean curvature flow. *Comm. Anal. Geom.* **10** (5) (2002), 1075–1113.

[V1] Viterbo, C., A new obstruction to embedding Lagrangian tori. *Invent. Math.* **100** (1990), 301–320.

[V2] Viterbo, C., Symplectic topology as the geometry of generating functions. *Math. Ann.* **292** (1992), 685–710.

[V3] Viterbo, C., On the uniqueness of generating Hamiltonian for continuous limits of Hamiltonians flows. Preprint, 2005; math.SG/0509179.

Department of Mathematics, University of Wisconsin, WI 53706, U.S.A.
and
Korea Institute for Advanced Study, Seoul, Korea
E-mail: oh@math.wisc.edu

Department of Mathematics, Kyoto University, Kitashirakawa, Kyoto, Japan
E-mail: fukaya@math.kyoto-u.ac.jp

Properly embedded minimal surfaces with finite topology

Antonio Ros*

Abstract. We present a synthesis of the situation as it now stands about the various moduli spaces of properly embedded minimal surfaces of finite topology in flat 3-manifolds. This family includes the case of minimal surfaces with finite total curvature in $\mathbb{R}^3$ as well as singly, doubly and triply periodic minimal surfaces.

Mathematics Subject Classification (2000). Primary 53A10; Secondary 53C42.

Keywords. Minimal surfaces, flux, least area.

1. Introduction

In these notes we will consider minimal surfaces Σ of finite topology which are properly embedded in $\mathbb{R}^3$ or in a complete flat 3-manifold $M = \mathbb{R}^3/\mathcal{G}$. For most of purposes, up to passing to a finite covering, we can assume that Σ is orientable and $\mathcal{G}$ is a cyclic group of rank 1, 2 or 3, which correspond to the singly, doubly and triply periodic cases, respectively. In the singly periodic case, $\mathcal{G}$ is generated by a screw motion (which in particular could be a translation). In the other cases, $\mathcal{G}$ consisting only on translations and M is either a flat 2-torus times $\mathbb{R}$, $T^2 \times \mathbb{R}$, or a flat 3-torus T^3.

We will focus on uniqueness and classification results. We will also emphasize those ideas and techniques which are (or could be) useful to understand the structure of moduli spaces of minimal surfaces. Although we have a large number of results in this area, several important questions about these moduli spaces remain unanswered.

There is also an interesting theory for the family of properly embedded minimal surfaces of finite genus and infinitely many ends, see for instance Meeks, Pérez and Ros [33] for structure results and Hauswirth and Pacard [13] for some recent examples. However we will not consider this situation in this paper. It is worth noticing that Colding and Minicozzi have proved that complete minimal surfaces of finite topology embedded in $\mathbb{R}^3$ are necessarily proper [3]. The same result holds in flat 3-manifolds, see [51].

Most of the surfaces we will consider have finite total curvature. They form an important and natural subclass. We refer the reader to the texts Pérez and Ros [50] and Hoffman and Karcher [15] and references therein, for more details about these surfaces.

*Partially supported by MCYT-FEDER research project MTM2004-02746.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

2. Geometry of the ends

An important achievement of the last years has been the complete understanding of the asymptotic geometry of these surfaces: the ends approximate simple model surfaces like the plane, the Catenoid or the Helicoid. Hoffman and Meeks [17] and Collin [5] showed that each end of a properly embedded minimal surface in Σ in $\mathbb{R}^3$ with finite topology and more than one end is asymptotic to either a plane or a Catenoid and the ends are all parallel (henceforth we will assume that these ends are horizontal). If Σ has just one end, then Meeks and Rosenberg [38] prove that Σ approaches to a Helicoid. The proof depends on recent results concerning limits of embedded minimal surfaces without curvature bounds, see Colding and Minicozzi [4], [40] and references therein.

In the periodic case, the geometry of the surface at infinity has been determined by Meeks and Rosenberg [35], [36], see Table 1 (note that when M is a 3-torus, we are just considering compact minimal surfaces). Assuming $M = \mathbb{R}^3/S_\theta$, where S_θ denotes a screw motion of angle θ and vertical axis, each end is asymptotic either to a plane (when $\theta \neq 0$, the plane must be horizontal), or to a flat vertical annulus like in the Scherk surface, see Figure 1 (this kind of end occurs only if θ is rational), or to the end of a vertical Helicoid.

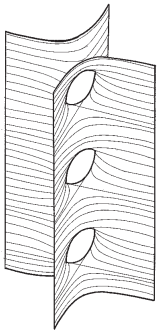


Figure 1. The singly periodic Scherk surface (seeing as a surface in the quotient space) has genus zero and four ends asymptotic to half-cylinders. It is a 1-parameter family of surfaces (the parameter is given by the angle between two consecutive wings).

If $M = T^2 \times \mathbb{R}$, then Σ has an even number of top (resp. bottom) ends and all of them are of Scherk type. The top ends are all parallel and the same holds for the bottom ones. Moreover, in case the top ends are not parallel to the bottom ends, then all the ends are vertical.

The above description of the ends has strong consequences on the geometry and the conformal structure of the minimal surface: Σ is conformally equivalent to a closed Riemann surface $\overline{\Sigma}$ with a finite number of punctures and the surface Σ can be described, by means of the Weierstrass representation, in terms of meromorphic data on the compactified surface $\overline{\Sigma}$.

Table 1. The admissible behaviour of the ends of a nonflat finite topology minimal surface. All the cases, but the first one, have finite total curvature. In the singly periodic case, all the ends of a given surface must be of the same type. The non-periodic helicoidal end is asymptotic to the Helicoid in $\mathbb{R}^3$, while the singly periodic one is asymptotic to that surface in $\mathbb{R}^3/S_\theta$.

periodicity	kind of ends
non-periodic	one helicoidal end
non-periodic	planar or catenoidal (more than one end)
singly-periodic	planar, helicoidal or Scherk ends
doubly-periodic	Scherk type

3. Minimal surfaces with finite topology in $\mathbb{R}^3$

Given integers $k \geq 0$ and $r \geq 1$, let $\mathcal{M}(k, r)$ be the moduli space of minimal surfaces $\Sigma \subset \mathbb{R}^3$ of genus k and r horizontal ends, properly embedded in $\mathbb{R}^3$. The simplest examples in this family can be characterized in terms of its topology.

Theorem 3.1. *The only properly embedded minimal surfaces of finite topology and genus zero in $\mathbb{R}^3$ are the Plane, the Catenoid and the Helicoid.*

If the surface has more than one end, the above result was proved by López and Ros [26]. In the one ended case this is a recent result of Meeks and Rosenberg [38]. The uniqueness of the Helicoid was a long standing problem which has been solved by using results of Colding and Minicozzi [4], [40].

Theorem 3.2. *The unique properly embedded minimal surface of finite topology in $\mathbb{R}^3$ with two ends is the Catenoid.*

Observe that in this characterization we prescribe only the number of ends, but not the genus. It was proved by Schoen [60] using the Alexandrov reflexion technique. A result in the same spirit has been obtained recently by Meeks and Wolf [39]: they prove that the singly periodic Scherk surface is characterized as the unique properly embedded minimal surface in $\mathbb{R}^3/T$ with four ends of Scherk type.

The first examples of higher genus where obtained by Costa [7] and Hoffman and Meeks [16] in the eighties. They constructed surfaces $\Sigma(k)$ of genus $k \geq 2$, two catenoidal ends and a middle planar end. The picture of the surface can be described as follows: each horizontal plane, other that $x_3 = 0$, meets the surface in a Jordan curve and $\Sigma \cap \{x_3 = 0\}$ consists on an equiangular system of $k + 1$ straight lines, see Figure 2.

These examples can be characterized as the ones of maximal symmetry in term of the genus of the surface.

Theorem 3.3. *Let Σ be a properly embedded minimal surface $\mathbb{R}^3$ with finite topology and positive genus. Then the symmetry group of Σ satisfies $|\mathrm{Sym}(\Sigma)| \leq$*

4(genus(Σ)+1). *Moreover, the equality holds if and only if* Σ *is one of the three-ended surfaces* $\Sigma(k)$ *constructed by Costa, Hoffman and Meeks.*

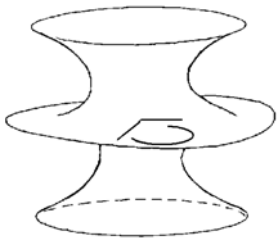


Figure 2. The Costa surface has genus one and three ends: the middle one is planar and the other are of catenoidal type. The surface has two vertical mirror planes and two horizontal reflection axes contained in the surface, but not in the mirror planes.

In the case Σ has three ends, the uniqueness in Theorem 3.3 was proved by Meeks and Hoffman [16]: the assumptions allows us to determine the symmetry group $\mathcal{G}$, the conformal structure of the surface, its picture in $\mathbb{R}^3$, up to a $\mathcal{G}$-invariant isotopy, and, finally, its Weierstrass data. This analysis has been used in several situations to produce new examples or to classify minimal surfaces with prescribed symmetry, see §1 in [1] for a general description of the method and concrete applications (in the final step we meet the so called *periods problem* which can be completely solved only in some cases). If the number of ends r is larger than 3, then the same kind of analysis shows that $r = 4$, $x_3 = 0$ is a mirror plane of the surface, $\Sigma \cap \{x_3 = 0\}$ consists of $k + 1$ Jordan curves with pairwise disjoint interior and $\Sigma \cap \{x_3 \geq 0\}$ is a surface of genus zero, 2 ends and $k + 1$ boundary components like in Figure 3, see Ros [54]. Finally we can prove that this surface does not exists by using the *vertical*

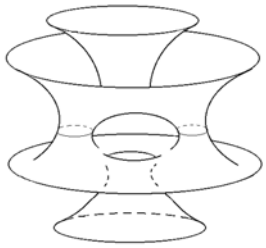


Figure 3. In $\mathbb{R}^3$, there are not properly embedded minimal surfaces of genus one, finitely many horizontal ends and an horizontal mirror plane. This can be shown by using the vertical flux deformation argument.

flux deformation argument, see [54] and Section 5 below. In fact, the surfaces exist as immersed surfaces, but they are not embedded. If $r = 1$, then the end is helicoidal and, therefore, the order of the symmetry group is at most 4. So this case is discarded.

The surface $\Sigma(k)$ admits a 1-parameter deformation by means of embedded minimal surfaces $\Sigma(k,t)$, $t \in \mathbb{R}$, with $\Sigma(k,0) = \Sigma(k)$ and three catenoidal ends for $t \neq 0$. Its symmetry group is generated by k vertical reflexion planes in equiangular position, see Hoffman and Karcher [15].

Costa [8] classified minimal tori with three punctures. This is the only positive genus case where the moduli space is completely known. The proof depends on the properties of elliptic functions.

Theorem 3.4. *Any properly embedded minimal surface in* $\mathbb{R}^3$ *with genus* 1 *and three ends lies in the family above,* $\mathcal{M}(1,3) = \{\Sigma(1,t)\}$.

For genus larger than 1, the above surfaces can be characterized in terms of its symmetries. The following classification theorem by Martin and Weber [30] extends previous results in [15], [25].

Theorem 3.5. *Let* Σ *be a properly embedded minimal surface in* $\mathbb{R}^3$ *with three ends and genus* $k \geq 2$. *If* $|\operatorname{Sym}(\Sigma)| \geq 2(k+1)$, *then* Σ *is one the of surfaces* $\Sigma(k,t)$.

A central and basic open problem concerning finite topology minimal surfaces in $\mathbb{R}^3$ is the following one, see Hoffman and Meeks [16].

Conjecture. The moduli space $\mathcal{M}(k,r)$ is empty for $r > k+2$.

In the case $k = 0$, this is has been proved in [26]. For higher genus, Meeks, Pérez and Ros [34] have proved that, given k, $\mathcal{M}(k,r)$ is empty for r large enough.

In the one-ended case, Hoffman, Weber and Wolf [18] have constructed a properly embedded minimal surface of genus one and one helicoidal end. No characterization of this example is known at the present. Meeks and Rosenberg [38] have proposed the following question (which they proved for $k = 0$).

Conjecture. For each $k = 0, 1, \ldots$, there is a unique properly embedded minimal surface in $\mathbb{R}^3$ of genus k and one end.

There is a large number of examples constructed by *desingularization*, see for instance Kapouleas [21] and Traizet [61] for the non-periodic case and Traizet and Weber [62], [64] for the periodic one.

There is another group of ideas (depending on conformal geometry tools as flat structures, Teichmüller theory, extremal length, …) which is well-adapted to prove existence, nonexistence and deformability results for minimal surfaces with prescribed symmetric-isotopy class, when the fundamental region of the surface is a disc bounded by mirror lines. The method has been developed by Weber and Wolf [65], [66], [67]. For instance, it can be shown by this method that a genus three surface, with the same symmetries than the Costa surface, and in the symmetric isotopy class of the surface in Figure 4 cannot be realized by a minimal surface, see [30].

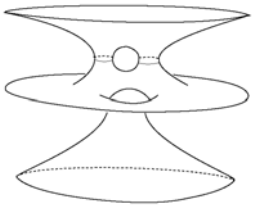


Figure 4. There are no properly embedded minimal surfaces with genus three, three ends, the same symmetries than the Costa surface and belonging to the symmetric-isotopy class of this figure. This can be shown by using flat structures and extremal length arguments.

4. The periodic case

There are a large number of examples of periodic minimal surfaces. As we are interested in those surfaces which have been characterized in some way, we will mention only a few of them.

Meeks [31] has classified the moduli space of minimal Klein bottles with a handle in flat 3-tori. These surfaces have (absolute) total curvature equal to 4π and correspond to the largest Euler characteristic among nonflat closed minimal surfaces in tori.

Theorem 4.1. *The space of closed minimal surfaces Σ in flat 3-tori with total curvature 4π (or, equivalently nonorientable surfaces with $\chi(\Sigma) = -2$) is parametrized by the family of antipodally invariant sets X in the 2-sphere with $\#(X) = 8$. Each X produces two surfaces, one and its conjugate. The surfaces Σ are all embedded.*

The next simplest situation to be understood is the genus 3 case, which contains in particular the orientable double covering of the Meeks surfaces above. There are genus 3 minimal surfaces, like the Schoen Gyroid [59], which cannot be obtained in this way. The following is a basic open question.

Problem. Give an explicit description of the moduli space of closed minimal surfaces of genus 3 in flat 3-tori.

Classical triply periodic minimal surfaces have a large symmetry group. Interestingly, several important examples have been constructed by crystallographers, see for instance Schoen [59], Fischer and Koch [11] and Lord and Mackay [27]. Many of these examples deserve a more exhaustive mathematical treatment, see Karcher [23] and Huff [19] for some results in this direction. To fix that idea, we focus now in a concrete question. Fischer and Koch [11] (see also Kawasaki [24]) have classified all the closed spatial polygons $\Gamma \subset \mathbb{R}^3$ which produce embedded triply periodic minimal surfaces by means of the following routine:

1) Construct a discoidal patch Δ by solving the Plateau problem for the boundary Γ.
2) After reflecting the patch Δ with respect to the edges of Γ, and so on, we get an embedded triply periodic minimal surface.

Most of the polygons Γ project monotonically onto a convex polygon in a plane. As a well-known consequence of the maximum principle, each of these Γ bound a unique minimal surface Δ. Therefore we deduce a number of uniqueness results for triply periodic minimal surfaces in terms of symmetry and topology. Eleven families of surfaces can be characterized in this way, see Fisher and Koch [11]. However there are four of these polygons, the ones named S, Y, $C(S)$ and $C(Y)$, which do not satisfy the convexity condition above, and so, the existence and uniqueness question remains to be clarified. In fact the surfaces Y and $C(S)$ can be realized, at least, by the D and P Schwarz surfaces, respectively (to do that we need to take on D and P patches larger than the usual ones) but other realizations cannot not be discarded at the moment.

Among noncompact periodic minimal surfaces which can be characterized in terms of its topology and symmetry, we have the following theorem, concerning a singly periodic version of Costa, Hoffman and Meeks surfaces, which combines existence results of Callahan, Hoffman and Meeks [1] with a uniqueness property by Martín and Rodriguez [29], [28].

Theorem 4.2. *Let $\Sigma \subset \mathbb{R}^3/S_\theta$ be a properly embedded minimal surface of genus $k \geq 2$ and two planar ends. Then $|\operatorname{Sym}(\Sigma)| \leq 4(k+1)$. Moreover, if the equality holds, then k is odd, $(k+1)\theta \in 4\pi\,\mathbb{Z}$ and Σ is one of the (translation invariant) surfaces constructed by Callahan, Hoffman and Meeks.*

It would be interesting and very useful to have a complete list of properly embedded minimal surfaces in flat three manifolds with small total curvature, or more generally, minimal surfaces with small total curvature modulo symmetries. As a first step, we propose the following more concrete question.

Problem. Classify properly embedded minimal surfaces in flat 3-manifolds with (absolute) total curvature smaller than or equal to 4π.

We have ten different types of compact flat 3-manifolds (some of them admit minimal surfaces of total curvature 2π).

5. Vertical flux

According to the *Weierstrass representation*, an orientable minimal surface in $\mathbb{R}^3$ can be represented by the data (Σ, g, ω), where Σ is a Riemann surface, g is a meromorphic map on Σ which corresponds (up to stereographic projection) to the Gauss map of the surface and ω is an holomorphic 1-form vanishing just at the poles of g (and the order of the zero being double of the order of the pole). The minimal surface is recovered as the immersion $\psi\colon \Sigma \to \mathbb{R}^3$ given by

$$\psi = \Re \int \left(\frac{1}{2}(1-g^2), \frac{i}{2}(1+g^2), g\right)\omega. \tag{1}$$

In order the immersion to be globally well-defined, we require that the real part of the periods of the above integral vanish. More generally, if the surface is periodic, the real part of the periods of (1) must be compatible with the prescribed periodicity. Several aspects of the geometry of the minimal surface are reflected in its Weierstrass representation.

If C is a closed curve on Σ, the *flux along the curve* is defined as the integral of the unit conormal vector, see Figure 5. This corresponds with the imaginary part of the periods of (1).

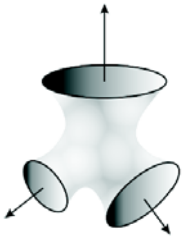


Figure 5. Flux of a minimal surface along its boundary components.

Among the simplest deformations of a minimal surface by minimal surfaces we have the *associated family*, given by $\Sigma_\theta = (\Sigma, g, e^{i\theta}\omega)$, $0 \leq \theta < 2\pi$, and the *vertical flux deformation* $\Sigma_\lambda = (\Sigma, \lambda g, \frac{1}{\lambda}\omega)$, $\lambda > 0$. In the first one the Gauss map and the induced metric are preserved. This deformation is globally well-defined if and only if all fluxes vanish. The second deformation fixes the third coordinate of the immersion and transforms the normal direction in a simple conformal way. It gives globally well-defined immersions if and only if the flux of any curve on Σ is vertical. An important difference between both deformations is that the first one consists on a compact family of surfaces while the second one is noncompact. After reparametrization and change of scale in $\mathbb{R}^3$, the part of the surface around a point where g has a zero converges, when λ goes to infinity, to the surface given by the Weierstrass data $(\mathbb{C}, z^n, a\,dz)$, $a \in \mathbb{C}^*$, which is not embedded. In a neighborhood of an end of the surface given by a punctured disc $0 < |z| < \varepsilon$ with meromorphic g and ω and $g(0) = 0$, Σ_λ converges to the minimal surface $(\mathbb{C}^*, z^n, a\,z^m dz)$, $a \in \mathbb{C}^*$. It can be checked easily that the only surfaces of this type which are embedded are the (vertical) Catenoid and the Helicoid. These surfaces correspond to the cases $n = 1, m = -2, a = 1$ and $m = n - 1, a = i$, respectively. Planar, catenoidal and helicoidal ends with vertical normal directions transform by the λ-deformation into ends of the same type (note that the flux at this ends is always vertical). Under suitable global assumptions this deformation gives strong restrictions on the geometry and the topology of minimal surfaces all of whose fluxes are vertical, see works of López, Pérez and Ros [26], [47], [54], [44]. In particular we have the following result.

Theorem 5.1. *A nonflat properly embedded minimal surface $\Sigma \subset \mathbb{R}^3$ of finite topology, horizontal ends and vertical flux is either a Catenoid or an Helicoid.*

As a minimal surface in the hypothesis of Theorem 3.1 has necessarily vertical flux, this theorem follows from the result above.

Now we explain briefly the proof of Theorem 5.1. First observe that the deformation Σ_λ, $\lambda > 0$, is well defined. It follows easily from the maximum principle for minimal surfaces that, if we start with an embedded nonflat triply-periodic minimal surface and we deform it continuously by triply-periodic minimal surfaces, embeddedness is preserved along the deformation. For general properly embedded minimal surfaces this fact is not generally true and it depends of the behaviour of the deformation at infinity. The maximum principle at infinity [37] and the behaviour of the vertical flux deformation at the ends allows to conclude that the surfaces Σ_λ are all embedded. Taking λ going to zero and infinity we conclude that no point on Σ has vertical normal vector. In the same way, we deduce that Σ has no planar ends and we obtain directly that Σ is homeomorphic either to a plane or a annulus. Now the theorem follows from [43], [38].

Some uniqueness results for singly periodic embedded minimal surfaces of finite topology, can be obtained by the arguments above:

i) there are not genus one minimal surfaces in $\mathbb{R}^3/S_\theta$, $\theta \neq 0$, with finitely many horizontal planar ends, see [47], and

ii) the only genus zero minimal surface in $\mathbb{R}^3/T$ with finitely many helicoidal ends is the Helicoid, [44].

Table 2. The surfaces in the first column are characterized as the unique embedded minimal surfaces satisfying the restrictions in the other columns. The last three results follow from the vertical flux deformation argument. The genus must be computed in the quotient surface.

surface	periodicity	genus	ends
Helicoid	none	0	one end
Catenoid	none	whatever	two ends
Catenoid	none	0	more than one
Helicoid	translation	0	helicoidal
none	screw $\theta \neq 0$	1	planar

6. Compactness and limit configurations

Consider a sequence $\{\Sigma_n\} \subset \mathcal{M}(k, r)$, where $\Sigma_n \subset \mathbb{R}^3$ is a properly embedded minimal surface of genus k and $r \geq 2$ horizontal ends. As the Gauss map $g_n \colon \overline{\Sigma}_n \to \overline{\mathbb{C}}$ is a meromorphic map of fixed degree, it converges up to a subsequence, to a family of nonconstant meromorphic maps $g_{\infty,1} \colon \overline{\Sigma}_{\infty,1} \to \overline{\mathbb{C}}, \ldots, g_{\infty,m} \colon \overline{\Sigma}_{\infty,m} \to \overline{\mathbb{C}}$, defined

over closed Riemann surfaces with $\text{degree}(g_{\infty,1})+\cdots+\text{degree}(g_{\infty,m}) = \text{degree}(g_n)$. For large n, one can see inside Σ_n large pieces of the surfaces $\overline{\Sigma}_{\infty,s}$ joined by regions with almost constant g_n.

A more careful analysis, see Ros [53], shows that each one of these meromorphic maps $g_{\infty,s}$ is the Gauss map of a properly embedded minimal surface $\Sigma_{\infty,s} \subset \mathbb{R}^3$ with horizontal ends, and that suitably chosen homothetic images of Σ_n converge to the different $\Sigma_{\infty,s}$. Moreover the regions joining these surfaces consist of *unbounded pieces* Ω_i, $i = 1, \ldots, r$ satisfying the following properties:

a) Each Ω_i contains exactly one end of Σ_n. So, the unbounded pieces are naturally ordered by their levels in $\mathbb{R}^3$,

b) the projection of Ω_i over the plane $\{x_3 = 0\}$ is one-to-one, and

c) the boundary of Ω_i consists of several convex Jordan curves in horizontal planes.

Therefore, for large n, the surface Σ_n looks like the one in Figure 6. As an example, the three-ended surfaces $\Sigma(k, t)$ described in §3 converge, when t goes to infinity, to a Catenoid between the level 1 and 2 and $k + 1$ Catenoids between levels 2 and 3.

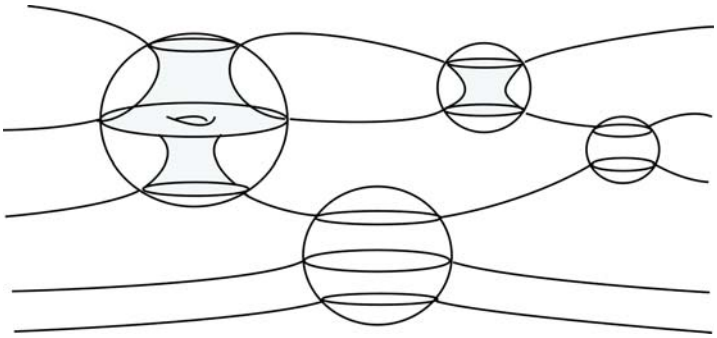


Figure 6. A sequence of minimal surfaces with fixed finite topology converges, up to a subsequence, to a union of surfaces with smaller topology joined by unbounded domains at different levels.

Working at the right scale, the limit can be seen as a horizontal plane of finite multiplicity with a finite number of marked points at different levels. Traizet [63] associates to this figure a number of horizontal forces (which correspond to rescaled limits of the fluxes of Σ_n). This *limit configuration* must be balanced in a natural sense.

We say that the moduli space $\mathcal{M}(k, r)$ is *compact (in the strong sense)* if any sequence $\{\Sigma_n\} \subset \mathcal{M}(k, r)$ converges (up to a subsequence) to a limit which consists of a single surface $\Sigma_\infty \in \mathcal{M}(k, r)$. In particular the moduli spaces $\mathcal{M}(k, 3)$, $k \geq 1$, are noncompact. The following theorem have been obtained by Ros [53] (for $r \geq 5$) and Traizet [63].

Theorem 6.1. *The moduli spaces* $\mathcal{M}(1, r)$, $r \geq 4$ *are compact* (*in the strong sense*).

In fact the compactness result can be extended (in a conditional but useful way) to higher genus. The space $\mathcal{M}(k, r)$ is compact for $r \geq g + 3$, assuming that the Hoffman–Meeks conjecture is true for genus smaller than k. This compactness can be seen as a first step in the proof of this conjecture. Thus Hoffman–Meeks conjecture will follow from the following one.

Conjecture. The moduli space $\mathcal{M}(k, r)$ is either empty or noncompact.

The compactness result in Ros [53] depends of the non existence of the piece described in Figure 7 in a limit of surfaces Σ_n with fixed topology. This piece consists

Figure 7. A sequence of minimal surfaces with fixed finite topology in $\mathbb{R}^3$ cannot converge to a limit which contains that piece, because the vertical flux argument gives a contradiction.

on an unbounded domain with just two catenoidal ends forming on it, one of positive and the other of negative logarithmic growth. The nonexistence of this piece is shown by using the vertical flux deformation argument.

The compactness theorem of Traizet [63] follows from the nonexistence in the limit of the surfaces Σ_n of a subsurface like the one in Figure 8. It is formed by

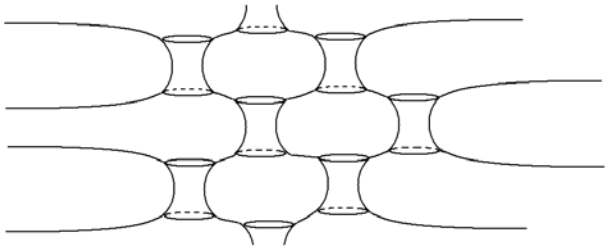


Figure 8. This picture cannot appear in a limit of a sequence of minimal surfaces in $\mathcal{M}(k, r)$. Otherwise, the limit configuration will be unbalanced.

several unbounded pieces at consecutive levels. Each one of this pieces is connected with the nearest ones by exactly two Catenoids forming. The top and the bottom unbounded pieces connect with the remaining part of the surface just by a catenoidal end forming. The reason why this piece cannot exist is because it is unbalanced.

The above ideas can be extended to the periodic case although the results have been explicitly stated only in some cases. Consider, for instance, a sequence $\{\Sigma_n\}$ of

genus one minimal surfaces with r horizontal planar ends, properly embedded in the flat three manifold $\mathbb{R}^3/T_n$, where T_n is a non-horizontal vector. It can be shown that the third coordinate has no critical points on $\overline{\Sigma}_n$ and so, each horizontal level curve is a Jordan curve (which might pass through one of the ends) and the Gauss map omits the vertical directions.

Up to a subsequence and suitable choice of scaling, $\{\Sigma_n\}$ converges either to a minimal surface $\Sigma_\infty \subset \mathbb{R}^3/T_\infty$ with the same topology than Σ_n or to a union of genus zero surfaces. Using the uniqueness results stated in Section 3 and further analysis, we can see that Σ_n approach to either r vertical Catenoids forming or to 2 vertical Helicoids forming, [32], see Figures 9 and 10. The second option is a simple example of the so called *parking garage* structure.

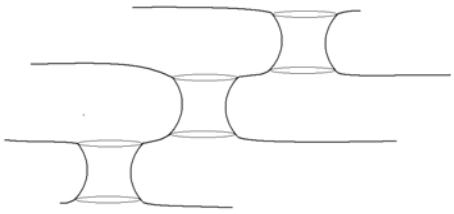


Figure 9. The Catenoid forming limit.

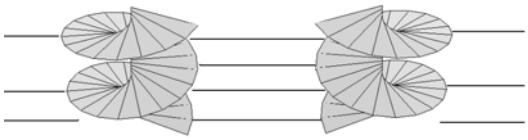


Figure 10. The Helicoid forming figure.

7. Smoothness of moduli spaces

Now we want to understand the structure of the moduli space $\mathcal{M}(k,r)$, $r \geq 2$, of finite topology minimal surfaces in $\mathbb{R}^3$ at its nonsingular points.

By expository reasons, we will consider in this section not $\mathcal{M}(k,r)$ but the subspace $\mathcal{M}'(k,r)$ of surfaces whose end logarithmic growths are all different (thus embeddedness is preserved by natural deformations in this space). Infinitesimal deformations of Σ in $\mathcal{M}'(k,r)$ are represented by *Jacobi functions*. These are smooth solutions $u\colon \Sigma \to \mathbb{R}$ of the equation $\Delta u + |\sigma|^2 u = 0$, where Δ is the Laplacian of the induced metric and $|\sigma|^2$ is the square length of the second fundamental form of the immersion. The functions u have at most logarithmic singularities at the ends of Σ,

which correspond to the fact that the growing of the catenoidal ends changes along the deformation. Denote by $\mathcal{J}(\Sigma)$ the space of these Jacobi functions. Using linear elliptic theory, it can be shown that $\dim \mathcal{J}(\Sigma) \geq r+3$, Pérez and Ros [48]. Moreover the subspace of nondegenerate surfaces $\mathcal{M}^*(k,r) = \{\Sigma \in \mathcal{M}'(k,r) : \dim \mathcal{J}(\Sigma) = r+3\}$ is a real analytic manifold, whose tangent space at a point Σ coincides with $\mathcal{J}(\Sigma)$. On this manifold we have the following additional structure: let $f\colon \mathcal{M}(k,r) \to \mathbb{R}^{2r}$ be the map which associates to a surface Σ the logarithmic growth of the asymptotic Catenoid and the *height of its neck* for each one of its ends. Then f induces a Lagrangian immersion of $\mathcal{M}^*(k,r)$ (modulo horizontal translations) in $\mathbb{R}^{2r-2}$, see [48]. We have been also able to compute the second fundamental form of this immersion, see Pérez and Ros [49].

If $\mathcal{B}(\Sigma)$ denotes the space of bounded Jacobi functions on Σ, then $\mathcal{B}(\Sigma)$ contains, at least the linear functions of the Gauss map (which correspond to the infinitesimal translations) and the function $\det(N, p, e_3)$, N, p and e_3 being the normal vector, the position vector and the vertical direction, respectively. This Jacobi function corresponds to the infinitesimal rotation of the surface around the vertical axis. It can be shown that if the above functions are the unique functions in $\mathcal{B}(\Sigma)$, then the surface is nondegenerate.

An interesting, and somewhat intriguing, fact is that bounded Jacobi functions in $\mathcal{B}(\Sigma)$ can be represented by branched conformal minimal immersions from $\overline{\Sigma} - B$ into $\mathbb{R}^3$, B being the ramification divisor of the Gauss map of Σ, whose Gauss map is the same than the one of Σ and whose ends have a bounded coordinate function, see Montiel and Ros [41] and Ejiri and Kotani [10]. Using this representation, Nayatani [42] has shown that the Costa, Hoffman and Meeks surfaces $\Sigma(k)$ are nondegenerate, for $k \leq 37$.

The local structure around a nondegenerate surface has been also considered in the periodic case and either planar or Scherk type ends [45], [14]. It would be interesting to clarify the nondegeneration condition and the smoothness properties of the moduli space of minimal surfaces in the case of helicoidal ends, both periodic and non-periodic.

Let $\mathcal{M}$ be the space of genus 3 embedded minimal surfaces in flat 3-tori. Then we can prove that there are some degenerate surfaces in $\mathcal{M}$, arguing as follows: assuming that any surface is nondegenerate, the subset $\mathcal{M}'$ of surfaces in $\mathcal{M}$ obtained as two sheeted coverings of non-orientable minimal surfaces of Euler characteristic -2 described in Theorem 4.1 is a union of connected components of $\mathcal{M}$. However, this is impossible as it is known that the Schwarz minimal surface $P \in \mathcal{M}'$, can be deformed to the Schoen Gyroid $G \in \mathcal{M} - \mathcal{M}'$.

An important open problem is to decide if a (generic) surface in $\mathcal{M}'(k,r)$ is nondegenerate. A related question is to give practical criteria which guarantee that a surface is nondegenerate. As an example, Montiel and Ros [41] proved that if all the branch values of the Gauss map (on the compactified surface $\overline{\Sigma}$) lie on a great circle, then the only bounded Jacobi functions are the linear functions of the Gauss map. In this way, we can prove the nondegeneracion of some surfaces, like finite coverings

of singly and doubly periodic Scherk surfaces, Riemann examples and Saddle towers constructed by Karcher [22].

8. Classification results

In this section we describe a strategy which has been used several times to classify surfaces in a moduli space. The first result of this type was proved by Meeks, Pérez and Ros [32] who classified genus zero properly embedded minimal surfaces in $\mathbb{R}^3$ with infinite symmetries. One of the key results in this classification is contained in the following theorem, see Figure 11.

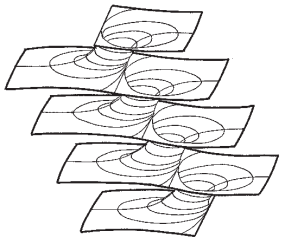


Figure 11. The Riemann minimal example.

Theorem 8.1. *A minimal surface Σ of genus 1 and r planar horizontal ends properly embedded in $\mathbb{R}^3/T$, T being a non horizontal vector, is a finite covering of one of the Riemann minimal examples.*

We will use this result to explain briefly how the method works. Consider the moduli space $\mathcal{M}$ of surfaces described in the theorem (r is fixed). As described in Section 6, any horizontal section C (whose level does not coincide with the level of an end) of a surface in $\mathcal{M}$ is a Jordan curve. The flux along C cannot be vertical, as the vertical flux deformation give a contradiction. Normalize the surfaces so that the third coordinate of the flux vector of C is 2π and define *the horizontal flux map* $F\colon \mathcal{M} \to \mathbb{R}^2 - \{0\}$ as the horizontal component of the flux along C (note that the flux of a curve around a planar end is zero and therefore the flux vector does not depend of the level we use to compute it). The following property follows from the results in Section 6.

I) *The map F is proper.*

In fact, the Catenoid forming picture has almost vertical flux which means that $F(\Sigma)$ converges to 0. In the Helicoid forming case the flux vector is almost horizontal and so $F(\Sigma)$ goes to ∞.

Denote by $\mathcal{R}$ the subspace of $\mathcal{M}$ formed by the coverings of the Riemann surfaces. The results in Section 7 imply that $\mathcal{R}$ is a smooth 2-dimensional manifold and moreover,

II) *$\mathcal{R}$ is an open an closed subset of $\mathcal{M}$, and $F\colon \mathcal{R} \to \mathbb{R}^2 - \{0\}$ is a diffeomorphism.*

We want to prove that F is an open map. This fact would follow, via the implicit function theorem, if we know that the surfaces in $\mathcal{M}$ are nondegenerate. However we do not have this property a priori. Instead we use the following complex variable theorem.

Theorem 8.2. *Let $f\colon \{z \in \mathbb{C}^m\colon |z| < \varepsilon\} \to \mathbb{C}^m$ be a holomorphic map with $f(0) = 0$. If 0 is an isolated point in $f^{-1}(0)$, then f is open around the origin.*

In our context we consider the space of *Weierstrass data* $\mathcal{W}$ which is a m-dimensional complex manifold. We also consider the *Period map* which consists on, both, the real and the imaginary parts of the periods which appear in the Weierstrass representation along a certain basis of the homology of the Riemann surface. A crucial point is that in this way we find a holomorphic map $P\colon \mathcal{W} \to \mathbb{C}^m$ between manifolds of the same dimension. This is a strong restriction which limitates the range of application of the whole method to problems where the involved surfaces are essentially (may be modulo symmetries) of genus zero.

In order a point Σ in $\mathcal{W}$ to define an immersed minimal surface, the real part of $P(\Sigma)$ must be zero. The imaginary part corresponds to the fluxes along the homology base and in our case this fluxes reduce to $F(\Sigma)$. Given a surface $\Sigma_0 \in \mathcal{M}$, the analytic subset $\mathcal{S} = \{\Sigma \in \mathcal{W} : P(\Sigma) = P(\Sigma_0)\}$ coincides, locally around Σ_0, with $\{\Sigma \in \mathcal{M} : F(\Sigma) = F(\Sigma_0)\}$ which is compact. After that we can deduce that Σ_0 is an isolated surface in $\mathcal{S}$, and then, using Theorem 8.2, we get that

III) *The map F is open.*

The final step is to prove that for some value of $\mathbb{R}^2 - \{0\}$ the pullback image of F consists only of Riemann examples. To do that we use an implicit function argument at the degenerate point of $\mathcal{W}$ given by the r Catenoids forming limit. Is can be shown that P extends holomorphically at that boundary point and has nonzero Jacobian. This proves that

IV) *Given $\Sigma \in \mathcal{M}$, if the length of $F(\Sigma)$ is small enough, then Σ is one of the Riemann examples.*

In Table 3 we have collected the different situations where the above strategy has been successfully applied. The first column contains the surfaces whose uniqueness have been shown. Lazard–Holly and Meeks [20] proved that the doubly periodic Scherk surface is the only genus zero minimal surface in $T^2 \times \mathbb{R}$. Pérez and Traizet [51] have proved recently that the Saddle towers constructed by Karcher [22] are the only examples of genus in $\mathbb{R}^3/T$ other than the Helicoid, and Rodriguez, Pérez and Traizet [46] characterized a 3-dimensional family of standard examples constructed by Karcher [22] and Meeks and Rosenberg [35], as the unique double periodic minimal surfaces of genus one and parallel Scherk type ends. Finally Meeks and Wolf [39]

Table 3. The (families of) surfaces in the first column have been characterized as the unique among surfaces satisfying the restrictions of the other columns. The topology we consider is the one of the quotient surface.

surfaces	periodicity	genus	ends
Riemann	singly	1	planar
Scherk	doubly	0	whatever
Saddle towers	single translation	0	Scherk
Standard examples	doubly	1	parallel Scherk
Scherk	single translation	whatever	four Scherk

have proved that the singly periodic Scherk surface is the unique surface in $\mathbb{R}^3/T$ with four Scherk ends. The proofs of these results follow formally the same steps than the one of the Riemann examples case, but the concrete arguments are more involved and several additional problems appear. In particular, the proof of the last result depends on a different group of ideas developed around the notion of orthodisks in Riemann surface theory. The following is a very natural problem which contains several of the results above.

Problem ([51]). Classify complete genus zero surfaces (of finite topology) embedded in complete flat 3-manifolds.

We remark that there exists a large family of genus zero surfaces with helicoidal ends in $\mathbb{R}^3/S_\theta$. These are called *twisted saddle towers* and were constructed by Karcher [22]. These surfaces are obtained as twisted deformations of the surfaces characterized in [51]. Other flat 3-manifolds may have genus zero minimal surfaces with ends of Scherk type.

9. Least area surfaces

Of course, the most natural class of minimal surfaces is the one of area minimizing surfaces. Although the plane is the unique least area surface in $\mathbb{R}^3$, there are nonplanar least area surfaces if we prescribe suitable symmetries. The complete description of area minimizing surfaces among surfaces satisfying a natural constraint, like to belong to a given symmetric isotopy class, is natural question. As a first goal, we propose the following more precise problem.

Problem. Classify area minimizing surfaces, modulo 2, in complete flat 3-manifolds.

Table 4. $\mathbb{Z}_2$-least area surfaces in the flat 3-manifolds obtained as a quotient of $\mathbb{R}^3$ by a translations group.

3-manifold	least area surfaces (mod 2)
$\mathbb{R}^3$	plane
$\mathbb{R}^3/T$	planar and 2π-Helicoid
$T^2 \times \mathbb{R}$, T^2 rectangular	planar and 2π-Scherk
$T^2 \times \mathbb{R}$, $T^2 \neq$ rectangular	planar
T^3	2-tori and $\chi(\Sigma) = -2$ (?)

It can be seen that any solution of this problem is either (a quotient of a) plane or an one-sided surface. In Table 4 we have listed the solutions of the problem for the quotients of $\mathbb{R}^3$ by translation groups. The result has been proved by Ros [55] and follows from the classification of complete stable surfaces in these 3-manifolds. The 2-sided case was solved by do Carmo and Peng [9], Fisher-Colbrie and Schoen [12] and Pogorelov [52], but the 1-sided remained open; for previous related results see Ross and Schoen [57], [58]. We prove that a complete stable minimal surface in $\mathbb{R}^3/T$ (resp. $T^2 \times \mathbb{R}$) is either planar or the nonorientable Helicoid (resp. Scherk surface) of total curvature 2π. We also prove that these surfaces are, both, area minimizing (mod 2). In the case of flat 3-tori, we prove that any stable nonflat closed minimal surface is a nonorientable surface with Euler characteristic equal to -2. Some surfaces of this topology are stable, like P and D Schwarz surfaces, see [56], but some other are unstable. However it is natural to hope that area minimizing surfaces in a flat 3-torus are flat 2-tori.

References

[1] Callahan, M., Hoffman, D., and Meeks, W. H., III, Embedded minimal surfaces with an infinite number of ends. *Invent. Math.* **96** (1989), 459–505.

[2] —, The structure of singly-periodic minimal surfaces. *Invent. Math.* **99** (1990), 455–481.

[3] Colding, T. H., and Minicozzi, W. P., II, The Calabi-Yau conjectures for embedded surfaces. math.DG/0404197.

[4] —, What are the shapes of embedded minimal surfaces and why? math.DG/0511740.

[5] Collin, P., Topologie et courbure des surfaces minimales proprement plongées de $\mathbf{R}^3$. *Ann. of Math.* (2) **145** (1997), 1–31.

[6] Cosín, C., and Ros, A., A Plateau problem at infinity for properly immersed minimal surfaces with finite total curvature. *Indiana Univ. Math. J.* **50** (2001), 847–878.

[7] Costa, C. J., Example of a complete minimal immersion in $\mathbb{R}^3$ of genus oneand three embedded ends. *Bull. Soc. Brasil. Math.* **15** (1984), 47–54.

[8] —, Classification of complete minimal surfaces in $\mathbb{R}^3$ with total curvature 12π. *Invent. Math.* **105** (1991) 273–303.

[9] do Carmo, M., and Peng, C. K., Stable complete minimal surfaces in $\mathbb{R}^3$ are planes. *Bull. Amer. Math. Soc.* (N.S.) **1** (1979), 903–906.

[10] Ejiri, N., and Kotani, M., Index and flat ends of minimal surfaces. *Tokyo J. Math.* **16** (1993), 37–48.

[11] Fischer, W., and Koch, E., Spanning minimal surfaces. *Philos. Trans. Roy. Soc. London Ser. A* **354** (1996), 2105–2142.

[12] Fischer-Colbrie, D., and Schoen, R., The structure of complete stable minimal surfaces in 3-manifolds of nonnegative scalar curvature. *Comm. Pure Appl. Math.* **33** (1980), 199–211.

[13] Hauswirth, L., and Pacard, F., Minimal surfaces of finite genus with two limit ends. Preprint.

[14] Hauswirth, L., and Traizet, M., The space of embedded doubly-periodic minimal surfaces. *Indiana Univ. Math. J.* **51** (2002), 1041–1080.

[15] Hoffman, D., and Karcher, H., Complete embedded minimal surfaces of finite total curvature. In *Geometry V: Minimal surfaces*, Encyclopaedia Math. Sci. 90, Springer-Verlag, Berlin 1997, 5–93.

[16] Hoffman, D., and Meeks, W. H., III, Embedded minimal surfaces of finite topology. *Ann. of Math.* **131** (1990), 1–34.

[17] —, The asymptotic behavior of properly embedded minimal surfaces of finite topology. *J. Amer. Math. Soc.* **2** (1989), 667–682.

[18] Hoffman, D., Weber, M., and Wolf, M., An Embedded Genus-One Helicoid. *Proc. Nat. Acad. Sci. U.S.A.* **102** (46) (2005), 16566–16568.

[19] Huff, R., Existence proofs of the C(H) and tC(P) surfaces. Preprint.

[20] Lazard-Holly, H., and Meeks, W. H., III, Classification of doubly-periodic minimal surfaces of genus zero. *Invent. Math.* **143** (2001), 1–27.

[21] Kapouleas, N., Complete embedded minimal surfaces of finite total curvature. *J. Differential Geom.* **47** (1997), 95–169.

[22] Karcher, H., Embedded minimal surfaces derived from Scherk's examples. *Manuscripta Math.* **62** (1988), 83–114.

[23] —, The triply periodic minimal surfaces of Alan Schoen and their constant mean curvature companions. *Manuscripta Math.* **64** (1989), 291–357.

[24] Kawasaki, T., Classification of spatial polygons that could possibly generate embedded triply periodic minimal surfaces. *Tokyo J. Math.* **26** (2003), 23–53.

[25] López, F. J., and Martin, F., Complete minimal surfaces in $\mathbb{R}^3$. *Publ. Mat.* **43** (1999), 341–449.

[26] López, F. J., and Ros, A., On embedded minimal surfaces of genus zero. *J. Differential Geom.* **33** (1991), 293–300.

[27] Lord, E. A., and Mackay, A. L., Periodic minimal surfaces of cubic symmetry. *Current Sci.* **85** (2003), 346–362.

[28] Martín, F., and Rodríguez, D., A characterization of the periodic Callahan-Hoffman-Meeks surfaces in terms of their symmetries. *Duke Math. J.* **89** (1997), 445–463.

[29] Martín, F., A note on the uniqueness of the periodic Callahan-Hoffman-Meeks surfaces in terms of their symmetries. *Geom. Dedicata* **86** (2001), 185—190.

[30] Martín, F., and Weber, M., Properly Embedded Minimal Surfaces with Three Ends. *Duke Math. J.* **107** (2001), 533–560.

[31] Meeks, W. H., III, The theory of triply periodic minimal surfaces. *Indiana Univ. Math. J.* **39** (1990), 877–936.

[32] Meeks, W. H., III, Pérez, J., and Ros, A., Uniqueness of the Riemann minimal examples. *Invent. Math.* **131** (1998), 107–132.

[33] —, The Geometry of Minimal Surfaces of Finite Genus I: Curvature Estimates and Quasiperiodicity. *J. Differential Geom.* **66** (2003), 1–45.

[34] —, The geometry of minimal surfaces of finite genus III: bounds on the topology and index of classical minimal surfaces. Preprint.

[35] Meeks, W. H., III, and Rosenberg, H., The global theory of doubly periodic minimal surfaces. *Invent. Math.* **97** (1989), 351–379.

[36] —, The geometry of periodic minimal surfaces. *Comment. Math. Helv.* **68** (1993), 538–578.

[37] —, The maximum principle at infinity for minimal surfaces in flat three manifolds. *Comment. Math. Helv.* **66** (1991), 263–278.

[38] —, The uniqueness of the helicoid and the asymptotic geometry of properly embedded minimal surfaces with finite topology. *Ann. of Math.*, to appear.

[39] Meeks, W. H., III, and Wolf, M., Minimal surfaces with the area growth of two planes; the case of infinite symmetry. Preprint

[40] Minicozzi, W. P., II, Embedded minimal surfaces. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 853–877.

[41] Montiel, S., and Ros, A., Schrödinger operators associated to a holomorphic map. In *Global differential geometry and global analysis* (ed. by D. Ferus, U. Pinkall, U. Simon and B. Wegner), Lecture Notes in Math. 1481, Springer-Verlag, Berlin 1991, 147–174.

[42] Nayatani, S., Morse index and Gauss map of complete minimal surfaces in Euclidean 3-space. *Comment. Math. Helv.* **68** (1993), 511–537.

[43] Osserman, R., *A survey of minimal surfaces*. 2nd ed., Dover, New York 1986.

[44] Pérez, J., A rigidity theorem for periodic minimal surfaces. *Comm. Anal. Geom.* **7** (1999), 95–104.

[45] —, On singly periodic minimal surfaces with planar ends. *Trans. Amer. Math. Soc.* **349** (1997), 2371–2389.

[46] Pérez, J., Rodríguez, M., and Traizet, M., The classification of doubly periodic minimal tori with parallel ends. *J. Differential Geom.* **69** (2005), 523–577.

[47] Pérez, J., and Ros, A., Some uniqueness and nonexistence theorems for embedded minimal surfaces. *Math. Ann.* **295** (1993), 513–525.

[48] —, The space of properly embedded minimal surfaces with finite total curvature. *Indiana Univ. Math. J.* **45** (1996), 177–204.

[49] —, The space of complete minimal surfaces with finite total curvature as lagrangian submanifold. *Trans. Amer. Math. Soc.* **351** (1999), 3935–3952.

[50] —, Properly embedded minimal surfaces with finite total curvature. In *The global theory of minimal surfaces in flat spaces* (ed. by G. P. Pirola), Lecture Notes in Math. 1775, Springer-Verlag, Berlin 2002, 15–66.

[51] Pérez, J., and Traizet, M., The classification of singly periodic minimal surfaces with genus zero and Scherk type ends. *Trans. Amer. Math. Soc.*, to appear.

[52] Pogorelov, A. V., On the stability of minimal surfaces. *Dokl. Akad. Nauk SSSR* **260** (1981), 293–295.

[53] Ros, A., Compactness of spaces of properly embedded minimal surfaces with finite total curvature. *Indiana Univ. Math. J.* **44** (1995), 139–152.

[54] —, Embedded minimal surfaces : forces, topology and symmetries. *Calc. Var. Partial Differential Equations* **4** (1996), 469–496.

[55] —, One-sided complete stable minimal surfaces. *J. Differential Geom.*, to appear.

[56] Ross, M., Schwarz' *P* and *D* surfaces are stable. *Differential Geom. Appl.* **2** (1992), 179–195.

[57] —, Complete nonorientable minimal surfaces in $\mathbf{R}^3$. *Comment. Math. Helv.* **67** (1992), 64–76.

[58] Ross, M., and Schoen, C., Stable quotients of periodic minimal surfaces. *Comm. Anal. Geom.* **2** (1994), 451–459.

[59] Schoen, A. H., Infinite periodic minimal surfaces without self-intersections. *NASA Technical Note No.* TN D-5541, 1970.

[60] Schoen, R., Uniqueness, symmetry and embeddedness of minimal surfaces. *J. Differential Geom.* **18** (1983), 791–809.

[61] Traizet, M., An embedded minimal surface with no symmetries. *J. Differential Geom.* **60** (2002), 103–153.

[62] —, Weierstrass representation of some simply-periodic minimal surfaces. *Ann. Global Anal. Geom.* **20** (2001), 77–101.

[63] —, A balancing condition for weak limits of minimal surfaces. *Comment. Math. Helv.* **79** (2004), 798–825.

[64] Traizet, M., and Weber, M., Hermite polynomials and helicoidal minimal surfaces. *Invent. Math.* **161** (2005), 113–149.

[65] Weber, M., On singly periodic minimal surfaces invariant under a translation. *Manuscripta Math.* **101** (2000), 125–142.

[66] Weber, M., and Wolf, M., Teichmüller theory and handle addition for minimal surfaces. *Ann. of Math.* **156** (2002), 713—795.

[67] Wolf, M., Flat structures, Teichmüller theory and handle addition for minimal surfaces. In *Global theory of minimal surfaces*, Clay Math. Proc. 2, Amer. Math. Soc., Providence, RI, 2005, 211–241.

Departamento de Geometría y Topología, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain

E-mail: aros@ugr.es

Applications of loop group factorization to geometric soliton equations

Chuu-Lian Terng*

Abstract. The 1-d Schrödinger flow on S^2, the Gauss–Codazzi equation for flat Lagrangian submanifolds in $\mathbb{R}^{2n}$, and the space-time monopole equation are all examples of geometric soliton equations. The linear systems with a spectral parameter (Lax pair) associated to these equations satisfy the reality condition associated to $\mathrm{SU}(n)$. In this article, we explain the method developed jointly with K. Uhlenbeck that uses various loop group factorizations to construct inverse scattering transforms, Bäcklund transformations, and solutions to Cauchy problems for these equations.

Mathematics Subject Classification (2000). 35Q55, 37K15, 37K35, 53C07.

Keywords. Soliton equations, space-time monopole equation, inverse scattering, Bäcklund transformations.

1. Introduction

A Hamiltonian system in $2n$-dimension is called *completely integrable* if it has n independent commuting Hamiltonians. By the Arnold–Liouville Theorem, such systems have *action-angle* variables that linearize the flow. The concept of completely integrability has been extended to soliton equations. These equation can be linearized using "scattering data", allowing one to use the Inverse Scattering method to solve the Cauchy problem with rapidly decaying initial data. Two model examples are the Korteweg–de Vries equation (KdV) and the non-linear Schrödinger equation (NLS). Soliton equations often arise naturally in differential geometry too. For example, the Gauss–Codazzi equations for surfaces in $\mathbb{R}^3$ with Gaussian curvature -1, isothermic surfaces in $\mathbb{R}^3$ [11], isometric immersions of space forms in space forms [15], [25], [24], Egoroff metrics, and flat Lagrangian submanifolds in $\mathbb{C}^n$ and $\mathbb{C}P^n$ [26], and the space-time monopole equation are soliton equations.

One of the key properties of a soliton equation is the existence of a Lax pair. A PDE for maps $q\colon \mathbb{R}^n \to \mathbb{R}^m$ admits a *Lax pair* if there exists a family of $\mathcal{G}$-connections θ_λ on $\mathbb{R}^n$, given in terms of q, such that the condition for θ_λ to be flat for all λ in an open subset of $\mathbb{C}$ is that q satisfy the PDE. The parameter λ is called the *spectral parameter*. For a solution q decaying at spatial infinity, we often can find a

*Research supported in part by NSF grant DMS- 052975.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

normalization so that there exists a unique parallel frame E_λ of θ_λ. Usually E_λ has two types of singularities for $\lambda \in \mathbb{C}P^1$; one type is a jump across a contour and the other type is a pole. We call the jump singularities of E_λ the *continuous scattering data for* q and the poles and residues of E_λ the *discrete scattering data for* q. The *scattering transform* maps a solution q to its scattering data S. A key feature of soliton PDEs is that the induced equations on the scattering data is linear, so it is easy to write down the scattering data of a solution at time t for a given initial data. The inverse scattering transform reconstructs q from the scattering data, i.e., it reconstructs E_λ from prescribed singularities. This is done for KdV in [17], for NLS in [36], [14], and for the n-wave equation in [36], [4], [5]. As a consequence, the Cauchy problem for these soliton equations can be solved via the inverse scattering transform.

The proof of the existence of the inverse scattering transform for soliton equations involves hard analysis and is difficult ([4]). However, if the Lax pair satisfies the $\mathrm{SU}(n)$-reality condition, then the frame $E_\lambda(x)$ has only one jumping circle plus pole singularities in the λ-sphere for each x, so the continuous scattering data is a loop into $\mathrm{SL}(n,\mathbb{C})$ for each x. In this case, we can use Pressley–Segal loop group factorization to construct the inverse scattering transform for the continuous scattering data (cf. [27]).

Bäcklund transformations (BTs) for surfaces in $\mathbb{R}^3$ with $K = -1$ arose from the study of line congruences in classical differential geometry. It associates to each surface in $\mathbb{R}^3$ with $K = -1$ a family of compatible systems of ordinary differential equations (ODEs) so that solutions of these ODE systems give rise to a family of new surfaces in $\mathbb{R}^3$ with $K = -1$. One can use line and sphere congruences to construct Bäcklund type transformations for many geometric problems in differential geometry (cf. [24]). Bäcklund transformations for soliton equations produce a new solution from a given one by adding discrete scattering data. These transformations can be obtained in a unified way from the following type of factorization: Let Γ_1, Γ_2 be disjoint subsets of S^2, and $g_i\colon S^2 \setminus \Gamma_i \to \mathrm{GL}(n,\mathbb{C})$ holomorphic for $i = 1, 2$. Factor $g_1 g_2 = \tilde{g}_2 \tilde{g}_1$ such that $\tilde{g}_i$ is holomorphic on $S^2 \setminus \Gamma_i$. This factorization can always be done when g_1 is rational and g_i satisfy the $\mathrm{SU}(n)$-reality condition, so global Bäcklund transformations exist for flows in the $\mathrm{SU}(n)$-hierarchy and for the space-time monopole equation with gauge group $\mathrm{SU}(n)$. Moreover, if the initial data q_0 has continuous scattering data S and discrete scattering data $\triangle$, then we can first use Pressley–Segal loop group factorization to construct a solution q whose scattering data is S, and then apply BTs to q to construct the solution $\tilde{q}$ with scattering data $S \cup \triangle$.

This paper is organized as follows: In Section 2, we outline the construction of the ZS-AKNS hierarchy of soliton equations associated to a complex simple Lie algebra $\mathcal{G}$, and review certain invariant submanifolds and restricted flows associated to involutions of $\mathcal{G}$. We give examples of PDEs in submanifold geometry that are soliton equations in Section 3. In Section 4, we give a brief review of Lax pairs associated to the space-time monopole equations. The direct scattering for soliton equations in the $\mathrm{SU}(n)$-hierarchy is given in Section 5, and direct scattering for space monopole equation is in Section 6. We use Pressley–Segal loop group factorization

to construct the inverse scattering transform for flows in the SU(n)-hierarchy and for the space-time monopole equation in Section 7 and 8 respectively. In Section 9, we use the Birkhoff factorization to construct local solutions for flows in the SU(n)-hierarchy. Finally, we discuss the constructions of Bäcklund transformations, pure solitons, and solutions with both continuous and discrete scattering data for flows in the SU(n)-hierarchy and for the space-time monopole equations in the last two sections.

Acknowledgment. The author thanks her long-time collaborator and good friend Karen Uhlenbeck. Much of this article concerns our joint project on the differential geometric aspects of soliton equations.

2. Soliton equations associated to simple Lie algebras

The method of constructing a hierarchy of $n \times n$ soliton flows developed by Zakharov–Shabat [39] and Ablowitz–Kaup–Newell–Segur [1] works equally well if we replace the algebra of $n \times n$ matrices by a semi-simple, complex Lie algebra $\mathcal{G}$ (cf. [18], [23], [27]).

The G-hierarchy. Let G be a complex, simple Lie group, $\mathcal{G}$ its Lie algebra, $\langle\ ,\ \rangle$ a non-degenerate, ad-invariant bilinear form on $\mathcal{G}$, $\mathcal{A}$ a maximal abelian subalgebra of $\mathcal{G}$, and $\mathcal{A}^\perp = \{\xi \in \mathcal{G} \mid \langle \xi, \mathcal{A}\rangle = 0\}$. Let $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp)$ denote the space of rapidly decaying maps from $\mathbb{R}$ to $\mathcal{A}^\perp$. Fix a regular element $a \in \mathcal{A}$ (i.e., the centralizer $\mathcal{G}_a = \mathcal{A}$). Then there is a unique family of $\mathcal{G}$-valued maps $Q_{b,j}(u)$ parametrized by $b \in \mathcal{A}$ and positive integer j satisfying the following recursive formula,

$$(Q_{b,j}(u))_x + [u, Q_{b,j}(u)] = [Q_{b,j+1}(u), a], \quad Q_{b,0}(u) = b, \tag{1}$$

and $\sum_{j=0}^{\infty} Q_{b,j}(u)\lambda^{-j}$ is conjugate to b as an asymptotic expansion at $\lambda = \infty$. In fact, $Q_{b,j}(u)$ is a polynomial in $u, \partial_x u, \ldots, \partial_x^{j-1} u$ (cf., [23], [27]). For $b \in \mathcal{A}$ and a positive integer j, the (b, j)-*flow* is the following evolution equation on $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp)$:

$$u_t = (Q_{b,j}(u))_x + [u, Q_{b,j}(u)] = [Q_{b,j+1}(u), a]. \tag{2}$$

The G-*hierarchy* is the collection of these (b, j)-flows.

The recursive formula (1) implies that u is a solution of the (b, j)-flow (2) if and only if

$$\theta_\lambda = (a\lambda + u)\, dx + (b\lambda^j + Q_{b,1}(u)\lambda^{j-1} + \cdots + Q_{b,j}(u))\, dt \tag{3}$$

is a flat $\mathcal{G}$-valued connection 1-form on the (x, t) plane for all $\lambda \in \mathbb{C}$. Here θ_λ is (left) flat, i.e., $d\theta_\lambda + \theta_\lambda \wedge \theta_\lambda = 0$. In other words, θ_λ is a Lax pair for the (b, j)-flow (2). Also θ_λ is flat is equivalent to

$$[\partial_x + a\lambda + u,\ \partial_t + b\lambda^j + Q_{b,1}(u)\lambda^{j-1} + \cdots + Q_{b,j}(u)] = 0.$$

The U-hierarchy. Let τ be an involution of G such that its differential at the identity e (still denoted by τ) is a conjugate linear involution on the complex Lie algebra $\mathcal{G}$, and U the fixed point set of τ. The Lie algebra $\mathcal{U}$ of U is a real form of $\mathcal{G}$. If $a, b \in \mathcal{U}$, then the (b, j)-flow in the G-hierarchy leaves $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{U})$-invariant (cf. [27]). The restriction of the flow (2) to $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{U})$ is the (b, j)-flow in the *U-hierarchy*. Since $Q_{b,j}(u)$ lies in $\mathcal{U}$, the Lax pair θ_λ defined by (3) is a $\mathcal{G}$-valued 1-form satisfying the *U-reality condition*:

$$\tau(\theta_{\bar\lambda}) = \theta_\lambda. \tag{4}$$

The U/K-hierarchy. Suppose U is the real form defined by the involution τ of G, and σ an involution of G such that $d\sigma_e$ is complex linear and $\sigma\tau = \tau\sigma$. Let K be the fixed point set of σ in U, $\mathcal{U}$ and $\mathcal{K}$ the Lie algebras of U and K respectively, and $\mathcal{P}$ the -1-eigenspace of $d\sigma_e$ on $\mathcal{U}$. Then U/K is a symmetric space, and $\mathcal{U} = \mathcal{K} + \mathcal{P}$. Let $\mathcal{A}$ be a maximal abelian subalgebra in $\mathcal{P}$. If $a, b \in \mathcal{A}$ and u in $\mathcal{A}^\perp \cap \mathcal{K}$, then the (b, j)-flow in the U-hierarchy leaves $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{K})$ invariant if j is odd, and is normal to $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{K})$ if j is even. The restriction of odd flows in the U-hierarchy to $\mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{K})$ is called the *U/K-hierarchy*. Moreover, θ_λ satisfies the *U/K-reality condition*

$$\tau(\theta_{\bar\lambda}) = \theta_\lambda, \quad \sigma(\theta_\lambda) = \theta_{-\lambda}.$$

Example 2.1. SL$(2, \mathbb{C})$-*hierarchy* (cf. [2]). Let $a = b = \operatorname{diag}(i, -i)$ and $\mathcal{A} = \mathbb{C}a$. Then

$$\mathcal{A}^\perp = \left\{ \begin{pmatrix} 0 & q \\ r & 0 \end{pmatrix} \mid q, r \in \mathbb{C} \right\},$$

$$Q_{a,1}(u) = u = \begin{pmatrix} 0 & q \\ r & 0 \end{pmatrix}, \quad Q_{a,2}(u) = \frac{i}{2} \begin{pmatrix} qr & \partial_x q \\ -\partial_x r & -qr \end{pmatrix},$$

$$Q_{a,3} = \frac{i}{4} \begin{pmatrix} q\partial_x r - r\partial_x q - \partial_x^2 q + 2q^2 r \\ -\partial_x^2 r + 2qr^2 - q\partial_x r + r\partial_x q \end{pmatrix}, \ldots.$$

The (a, j)-flows, $j = 1, 2, 3$, in the SL$(2, \mathbb{C})$-hierarchy are:

$$\begin{aligned} &\partial_t q = \partial_x q, && \partial_t r = \partial_x r, \\ &\partial_t q = \frac{i}{2}(\partial_x^2 q - 2q^2 r), && \partial_t r = -\frac{i}{2}(\partial_x^2 r - 2qr^2), \\ &\partial_t q = \frac{1}{4}(-\partial_x^3 q + 6qr\partial_x q), && \partial_t r = \frac{1}{4}(-\partial_x^3 r + 6qr\partial_x r). \end{aligned}$$

Let τ be the involution of $sl(2, \mathbb{C})$ defined by $\tau(\xi) = -\bar\xi^t$. Then the fixed point set of τ is the real form $\mathcal{U} = su(2)$ and

$$\mathcal{A}^\perp \cap \mathcal{U} = \left\{ \begin{pmatrix} 0 & q \\ -\bar q & 0 \end{pmatrix} \mid q \in \mathbb{R} \right\}.$$

So the SU(2)-hierarchy is the restriction of the SL(2, $\mathbb{C}$)-hierarchy to the subspace $r = -\bar{q}$. The second flow in the SU(2)-hierarchy is the NLS $\partial_t q = \frac{i}{2}(\partial_x^2 q + 2|q|^2 q)$.

Let $\sigma(\xi) = -(\xi^t)$. Then $\sigma\tau = \tau\sigma$ and the corresponding symmetric space is SU(2)/ SO(2). Note that $u \in \mathcal{S}(\mathbb{R}, \mathcal{A}^\perp \cap \mathcal{K})$ means $q = -r$ is real. The third flow in the SU(2)/ SO(2)-hierarchy is the mKdV equation $q_t = -\frac{1}{4}(q_{xxx} + 6q^2 q_x)$.

The U-system. Let $\mathcal{U}$ be the real form of $\mathcal{G}$ defined by the involution τ, $\mathcal{A}$ a maximal abelian subalgebra of $\mathcal{U}$, and $a_1, \ldots, a_n$ a basis of $\mathcal{A}$. The U-system is the following PDE for $v\colon \mathbb{R}^n \to \mathcal{A}^\perp$:

$$[a_j, \partial_{x_i} v] - [a_i, \partial_{x_j} v] + [[a_i, v], [a_j, v]] = 0, \quad i \neq j. \tag{5}$$

It has a Lax pair

$$\theta_\lambda = \sum_{i=1}^{n} (a_i \lambda + [a_i, v])\, dx_i. \tag{6}$$

This Lax pair satisfies the *U-reality condition* $\theta_\lambda = \tau(\theta_{\bar{\lambda}})$.

The U/K-system. Let $\tau, \sigma, U, K, \mathcal{P}, \mathcal{A}$ be as in the U/K-hierarchy, and $a_1, \ldots, a_n$ a basis of $\mathcal{A}$. The U/K-system is the restriction of (5) to the space of $v\colon \mathbb{R}^n \to \mathcal{A}^\perp \cap \mathcal{P}$. Since $a_i \in \mathcal{P}$ and $[a_i, v] \in \mathcal{K}$, its Lax pair $\theta_\lambda = \sum_{i=1}^n (a_i\lambda + [a_i, v])\, dx_i$ satisfies the *U/K-reality condition*.

The frame of a Lax pair. Suppose we are given a family of flat $\mathcal{G}$-valued connections $\theta_\lambda = \sum_{i=1}^n P_i(x, \lambda) dx_i$ on $\mathbb{R}^n$. Then we call $E(x, \lambda)$ *a frame* of θ_λ if $E^{-1}\partial_{x_i} E = P_i$ for all $1 \leq i \leq n$.

Proposition 2.2. *Let G, τ, σ, U and K be as above, and E_λ the frame of θ_λ such that $E_\lambda(0) = \mathrm{I}$.*

1. *If θ_λ satisfies the U-reality condition, then E_λ satisfies the U-reality condition $\tau(E_{\bar{\lambda}}) = E_\lambda$,*

2. *If θ_λ satisfies the U/K-reality condition, then E_λ satisfies the U/K-reality condition $\tau(E_{\bar{\lambda}}) = E_\lambda$, $\sigma(E_\lambda) = E_{-\lambda}$.*

3. Soliton equations in submanifold geometry

Since the Gauss–Codazzi equations for submanifolds in space forms are equivalent to the flatness of certain connections, it is not surprising that many PDEs in submanifold geometry turns out to be soliton equations. We give some examples below:

Example 3.1 (Vortex filament equation, Schrödinger flow on S^2, and the NLS). In 1906, da Rios modeled the movement of a thin vortex in a viscous fluid by the motion of a curve propagating in $\mathbb{R}^3$ by

$$\partial_t \gamma = \partial_x \gamma \times \partial_x^2 \gamma. \tag{7}$$

If γ is a solution of (7), then

$$\partial_t \langle \partial_x \gamma, \partial_x \gamma \rangle = 2\langle \partial_x \partial_t \gamma, \partial_x \gamma \rangle = 2\langle \partial_x (\partial_x \gamma \times \partial_x^2 \gamma), \partial_x \gamma \rangle = 0.$$

So (7) preserves arc-length. Hence we may assume that a solution $\gamma(x, t)$ of (7) satisfying $\|\partial_x \gamma\| = 1$. It is known that there exists a parallel normal frame $(v_1, v_2)(\cdot, t)$ for each curve $\gamma(\cdot, t)$ such that $q = k_1 + ik_2$ is a solution of the NLS, where k_1 and k_2 are the principal curvatures of γ along v_1 and v_2 respectively.

Let $\mathcal{E}$ denote the energy functional on the space of paths on S^2, and J the complex structure on S^2 (if we view $S^2 \subset \mathbb{R}^3$, then $J_u(v) = u \times v$). The Schrödinger flow on S^2 is

$$u_t = J_u(\nabla \mathcal{E}(u)) = u \times u_{xx}.$$

If γ is a solution of (7), then $u = \gamma_x$ is a solution of the Schrödinger flow on S^2 ([14], [28]).

Example 3.2 (Isothermic surfaces in $\mathbb{R}^3$). A parametrized surface $f(x, y) \in \mathbb{R}^3$ is *isothermal* if (x, y) is a conformal line of curvature coordinate system, i.e., the two fundamental forms are of the form

$$\mathrm{I} = e^{2u}(dx_1^2 + dx_2^2), \quad \mathrm{II} = e^u(r_1 dx_1^2 + r_2 dx_2^2).$$

The Gauss–Codazzi equation is the $\frac{O(4,1)}{O(3)\times O(1,1)}$-system (cf. [11], [10], [9]).

Example 3.3 (Local isometric immersions of $N^n(c)$ in $N^{2n}(c)$, [25]). Let $N^m(c)$ denote the n-dimensional space form of constant sectional curvature c. The normal bundle of a submanifold M in $N^m(c)$ is *flat* if its induced normal connection is flat, and is *non-degenerate* if the dimension of $\{A_v \mid v \in \nu(M)_p\}$ is equal to $\operatorname{codim}(M)$. Here A_v is the shape operator along normal vector v. It is proved in [25] that if M^n is a submanifold of $N^{2n}(c)$ with constant sectional curvature c and its normal bundle $\nu(M)$ is flat and non-degenerate, then there exists a local orthogonal coordinate system $(x_1, \ldots, x_n)$ on M and parallel normal frame $e_{n+1}, \ldots, e_{2n}$ such that

$$\mathrm{I} = \sum_{i=1}^n b_i^2 dx_i^2, \quad \mathrm{II} = \sum_{j=1}^n a_{ji} b_i dx_i^2 e_{n+j}. \tag{8}$$

Moreover, the Levi-Civita connection 1-form for I is $w = \delta F - F^t \delta$, where $F = (f_{ij})$, $f_{ij} = \frac{\partial_{x_j} b_i}{b_j}$ if $i \neq j$, $f_{ii} = 0$ for all $1 \leq i \leq n$, and $\delta = \operatorname{diag}(dx_1, \ldots, dx_n)$. The Gauss–Codazzi equation for the local isometric immersion becomes an equation for F, which is the $\frac{O(2n)}{O(n)\times O(n)}$-system if $c = 0$, the $\frac{O(2n+1)}{O(n+1)\times O(n)}$-system if $c = 1$, and the $\frac{O(2n,1)}{O(n)\times O(n,1)}$-system if $c = -1$.

Example 3.4 (Egoroff metrics and the $\frac{U(n)}{O(n)}$-system). A local orthogonal system $(x_1, \ldots, x_n)$ of $\mathbb{R}^n$ is *Egoroff* if the flat Euclidean metric ds^2 written in this coordinate system is of the form

$$ds^2 = \sum_{j=1}^{n} \partial_{x_i}\phi \, dx_i^2$$

for some smooth function ϕ. Then $F = (f_{ij})$ is a solution of the $\frac{U(n)}{O(n)}$-system, where $f_{ij} = \frac{\partial_{x_i}\partial_{x_j}\phi}{2\sqrt{\partial_{x_i}\phi\partial_{x_j}\phi}}$ if $i \neq j$ and $f_{ii} = 0$ for $1 \leq i \leq n$. Conversely, given a solution $F = (f_{ij})\colon \mathbb{R}^n \to V_n$ of the $\frac{U(n)}{O(n)}$-system, the first order system

$$\partial_{x_j} b_i = f_{ij} b_j, \quad i \neq j \tag{9}$$

is solvable for $b_1, \ldots, b_n$, and solutions are locally defined and depend on n functions of one variable. Moreover, since $f_{ij} = f_{ji}$, $\sum_{i=1}^{n} b_i^2 dx_i$ is closed, hence locally there exists a smooth function ϕ such that $b_i^2 = \partial_{x_i}\phi$ for $1 \leq i \leq n$.

Although we can construct global solutions F for the $U(n)/O(n)$-system, it is not clear whether there exist global solutions b_i of (9) such that $b_i > 0$ and the metric $ds^2 = \sum_{i=1}^{n} b_i^2$ is complete. This is also the case for isometric immersions of $N^n(c)$ in $N^{2n}(c)$ and for the next example.

Example 3.5 (Flat Lagrangian submanifolds in $\mathbb{R}^{2n}$). As seen in Example 3.3, the Gauss–Codazzi equation for local isometric immersions of $\mathbb{R}^n$ into $\mathbb{R}^{2n}$ with flat and non-degenerate normal bundle is the $\frac{O(2n)}{O(n)\times O(n)}$-system. These immersions are Lagrangian if and only if F is symmetric and F is a solution of the $\frac{U(n)}{O(n)}$-system.

4. The space-time monopole equation

For flows in the SU(n)-hierarchy, we have been using left flat connections $\theta = \sum_{i=1}^{n} A_i dx_i$, i.e., $d\theta + \theta \wedge \theta = 0$ or equivalently, $[\partial_{x_i} + A_i, \partial_{x_j} + A_j] = 0$ for all $i \neq j$. But for space-time monopole equations, it is more customary to use right flat connections, i.e., $d\theta - \theta \wedge \theta = 0$, or equivalently, $[\partial_{x_i} - A_i, \partial_{x_j} - A_j] = 0$ for all $i \neq j$.

The curvature of a $su(n)$-valued connection 1-form $A = \sum_{i=1}^{4} A_i(x) dx_i$ is $F_A = \sum_{i<j} F_{ij}\, dx_i \wedge dx_j$. where

$$F_{ij} = [\partial_{x_i} - A_i, \partial_{x_j} - A_j] = \partial_{x_j} A_i - \partial_{x_i} A_j + [A_i, A_j].$$

The connection A is anti self-dual Yang–Mills (ASDYM) if

$$F_A = - * F_A,$$

where $*$ is the Hodge star operator with respect to the metric $dx_1^2+dx_2^2-dx_3^2-dx_4^2$.

Set $z = x_1 + ix_2$, $w = x_3 + ix_4$, $\nabla_z = \frac{1}{2}(\nabla_1 - i\nabla_2) = \frac{\partial}{\partial z} - A_z$, $\nabla_{\bar z} = \frac{1}{2}(\nabla_1 + i\nabla_2) = \frac{\partial}{\partial \bar z} - A_{\bar z}$, and $\nabla_w, \nabla_{\bar w}$ similarly. Since $A_i \in u(n)$, $A_{\bar z} = -A_z^*$ and $A_{\bar w} = -A_w^*$. Then (cf. [8], [21]) A is ASDYM if and only if

$$[\nabla_{\bar w} + \mu\nabla_z,\ \nabla_w + \mu^{-1}\nabla_{\bar z}] = 0. \tag{10}$$

holds for all $\mu \in \mathbb{C} \setminus \{0\}$.

If we assume that the ASDYM connection A is independent of x_4, and set $x = x_1$, $x_2 = y$, and $x_3 = t$, then $A_w = \frac{1}{2}(A_t - i\phi)$ and $A_{\bar w} = \frac{1}{2}(A_t + i\phi)$, where $\phi = A_{x_4}$ is the Higgs field, $A = A_t\,dt + A_z\,dz + A_{\bar z}\,d\bar z$ is a connection 1-form on $\mathbb{R}^{2,1}$. Then (A, ϕ) satisfies the *space-time monopole equation*

$$D_A\phi = *F_A,$$

where $*$ is the Hodge star operator with respect to the metric $dx^2 + dy^2 - dt^2$. It has a Lax pair induced from (10):

$$\left[\frac{1}{2}\nabla_t - \frac{i\phi}{2} + \mu\nabla_z,\ \ \frac{1}{2}\nabla_t + \frac{i\phi}{2} + \mu^{-1}\nabla_{\bar z}\right] = 0. \tag{11}$$

Set

$$D_1(\mu) = \frac{1}{2}\nabla_t - \frac{i\phi}{2} + \mu\nabla_z, \quad D_2(\mu) = \frac{1}{2}\nabla_t + \frac{i\phi}{2} + \mu^{-1}\nabla_{\bar z},$$
$$\begin{cases} P_1(\mu) = D_1(\mu) - D_2(\mu) = \frac{\mu-\mu^{-1}}{2}\nabla_x - i\frac{\mu+\mu^{-1}}{2}\nabla_y - i\phi, \\ P_2(\mu) = D_1(\mu) + D_2(\mu) = \nabla_t + \mu\nabla_z + \mu^{-1}\nabla_{\bar z}. \end{cases}$$

So (11) is equivalent to

$$\left[\frac{\mu-\mu^{-1}}{2}\nabla_x - \frac{i(\mu+\mu^{-1})}{2}\nabla_y - i\phi,\ \nabla_t + \mu\nabla_z + \mu^{-1}\nabla_{\bar z}\right] = 0. \tag{12}$$

Note that the first operator is a linear operator in space variables. This is the Lax pair we use to construct monopoles with continuous scattering data.

We need an equivalent form of the Lax pair to construct soliton monopoles. First we make a change of coordinates and spectral parameter:

$$\xi = \frac{t+x}{2}, \quad \eta = \frac{t-x}{2}, \quad \mu = \frac{\tau - i}{\tau + i}.$$

A direct computation shows that

$$L_1 = (\tau+i)D_1(\mu) + (\tau-i)D_2(\mu) = \tau\nabla_\xi - \nabla_y + \phi,$$
$$L_2 = \frac{1}{i}((\tau+i)D_1(\mu) - (\tau-i)D_2(\mu)) = \tau(\nabla_y + \phi) - \nabla_\eta,$$

so $[\tau\nabla_\xi - \nabla_y + \phi,\ \tau(\nabla_y + \phi) - \nabla_\eta] = 0$. Change spectral parameter again by $\lambda = \tau^{-1}$ to get

$$[\lambda(\nabla_y - \phi) - \nabla_\xi,\ \lambda\nabla_\eta - \nabla_y - \phi] = 0. \tag{13}$$

This is the Lax pair we use to construct Bäcklund transformations and solitons for the monopole equation. So we have

Proposition 4.1. *The following statements are equivalent for* (A, ϕ)*:*

1. (A, ϕ) *is a solution of the space-time monopole equation,*
2. (11) *holds for all* $\mu \in \mathbb{C} \setminus \{0\}$*,*
3. (12) *holds* $\mu \in \mathbb{C} \setminus \{0\}$*,*
4. (13) *holds for all* $\lambda \in \mathbb{C}$*,*
5. *there exists* $E_\mu(x, y, t)$ *such that*

$$\begin{cases} \left(\frac{\mu-\mu^{-1}}{2}\,\partial_x - \frac{i(\mu+\mu^{-1})}{2}\,\partial_y\right)E_\mu = \left(\frac{\mu-\mu^{-1}}{2}\,A_x - \frac{i(\mu+\mu^{-1})}{2}\,A_y + i\phi\right)E_\mu, \\ (\partial_t + \mu\partial_z + \mu^{-1}\partial_{\bar z})E_\mu = (A_t + \mu A_z + \mu^{-1}A_{\bar z})E_\mu, \\ E^*_{\bar\mu^{-1}}E_\mu = \mathrm{I}, \end{cases} \tag{14}$$

6. *there exists* $\psi_\lambda(x, y, t)$ *such that*

$$\begin{cases} (\lambda\partial_y - \partial_\xi)\psi_\lambda = (\lambda(A_y + \phi) - A_\xi)\psi_\lambda, \\ (\lambda\partial_\eta - \partial_y)\psi_\lambda = (\lambda A_\eta - A_y + \phi)\psi_\lambda, \\ \psi^*_{\bar\lambda}\psi_\lambda = \mathrm{I}, \end{cases} \tag{15}$$

7. $E_\mu(x, y, t)$ *is a solution of* (14) *if and only if*

$$\psi_\lambda(x, y, t) = E_{\frac{1-i\lambda}{1+i\lambda}}(x, y, t)$$

is a solution of (15)*.*

We call solutions of (14) and (15) *frames* of the monopole (A, ϕ). But frames are not unique. In fact, if ψ_λ is a solution of (15) and ϕ_λ satisfies

$$(\lambda\partial_y - \partial_\xi)\phi_\lambda = (\lambda\partial_\eta - \partial_y)\phi_\lambda = 0, \quad \phi^*_{\bar\lambda}\phi_\lambda = \mathrm{I}, \tag{16}$$

then $\psi_\lambda\phi_\lambda$ is also a solution of (15). Moreover, given any meromorphic map $h\colon \mathbb{C} \to \mathrm{GL}(n, \mathbb{C})$ that satisfies $h(\bar\lambda)^* h(\lambda) = \mathrm{I}$, then $\phi_\lambda(x, y, t) = h(y + \lambda\xi + \lambda^{-1}\eta)$ is a solution of (16). However, if (A, ϕ) is rapidly decaying in spatial variables, then we can choose normalizations (boundary conditions at infinity) so that there is a unique frame satisfying the normalization.

5. Direct scattering for flows in the SU(n)-hierarchy

Let $V_n = \{(\xi_{ij}) \in su(n) \mid \xi_{ii} = 0 \ \forall\ 1 \le i \le n\}$. The phase space of evolution equations in the SU(n)-hierarchy is the set $\mathcal{S}(\mathbb{R}, V_n)$ of all smooth $u\colon \mathbb{R} \to V_n$ that are rapidly decaying.

Recall that u is a solution of the (b, j)-flow in the SU(n)-hierarchy if and only if

$$\begin{cases} \psi_\lambda^{-1}\partial_x\psi_\lambda = a\lambda + u, \\ \psi_\lambda^{-1}\partial_t\psi_\lambda = b\lambda^j + Q_{b,1}(u)\lambda^{j-1} + \cdots + Q_{b,j}(u), \\ \psi_{\bar\lambda}^{*}\psi_\lambda = \mathrm{I} \end{cases} \tag{17}$$

is solvable. Since u decays in x, it is natural to study solutions of the first linear operator in (17) of the form $e^{a\lambda x}m(x,\lambda)$. The direct scattering refers to the study of singularities of $m(x,\lambda)$ in spectral parameter λ. This was done by Beals and Coifman:

Theorem 5.1 ([4]). *If $u \in \mathcal{S}(\mathbb{R}, V_n)$, then there exist a bounded discrete subset $\triangle_u$ of $\mathbb{C}\setminus\mathbb{R}$ and a smooth map $m\colon \mathbb{R}\times\mathbb{C}\setminus(\mathbb{R}\cup\triangle_u) \to \mathrm{GL}(n,\mathbb{C})$ such that*

1. *$\psi(x,\lambda) = e^{a\lambda x}m(x,\lambda)$ satisfies $d_x\psi = \psi(a\lambda + u)$,*

(i) *$m(x,\bar\lambda)^* m(x,\lambda) = \mathrm{I}$ and $\lim_{x\to-\infty} m(x,\lambda) = \mathrm{I}$,*

(ii) *$m(x,\lambda)$ is holomorphic for $\lambda \in \mathbb{C}\setminus(\mathbb{R}\cup\triangle_u)$, has poles at points in $\triangle_u$, and $m_\pm(x,r) = \lim_{s\to 0^\pm} m(x, r+is)$ is smooth,*

(iii) *m has an asymptotic expansion at $\lambda = \infty$:*

$$m(x,\lambda) \sim \mathrm{I} + m_1(x)\lambda^{-1} + m_2(x)\lambda^{-2} + \cdots .$$

Moreover,

1. *there is an open dense subset $\mathcal{S}_0(\mathbb{R}, V_n)$ of $\mathcal{S}(\mathbb{R}, V_n)$ such that $\triangle_u$ is a finite set for $u \in \mathcal{S}_0(\mathbb{R}, V_n)$,*
2. *if the L^1-norm of u is less than 1, then $m(x,\lambda)$ is holomorphic in $\lambda \in \mathbb{C}\setminus\mathbb{R}$, i.e., $\triangle_u$ is empty,*
3. *set $S(x,r) = m_+(x,r)m_-(x,r)^{-1}$, then $S^* = S$ and $S(x,r) - \mathrm{I}$ is rapidly decaying in r.*

The function m in the above theorem is called the *reduced wave function for the operator* $d_x + a\lambda + u$, the poles and residues of m are called the *discrete scattering data*, and the jump S is called the *continuous scattering data* of $d_x + a\lambda + u$.

Theorem 5.2 ([4], [5]). *Let u be a solution of the (b, j)-flow (2) in the* SU(n)*-hierarchy such that $u(\cdot,t) \in \mathcal{S}(\mathbb{R}, V_n)$, $m(\cdot,t,\cdot)$ and $S(\cdot,t,\cdot)$ the reduced wave function and the continuous scattering data for $d_x + a\lambda + u(\cdot,t)$ respectively for each t. Set $\psi(x,t,\lambda) = e^{a\lambda x + b\lambda^j t}m(x,t,\lambda)$. Then:*

1. *ψ is a solution of* (17)*,*
2. $\begin{cases}\partial_x S = [S, ar], \\ \partial_t S = [S, br^j].\end{cases}$
   In particular one has $S(x,t,r) = e^{-(arx+br^j t)}s_0(r)e^{arx+br^j t}$ *for some mapping* $s_0 : \mathbb{R} \to \mathrm{GL}(n,\mathbb{C})$ *such that* $s_0^* = s_0$ *and* $s_0 - \mathrm{I}$ *is rapidly decaying.*
3. *If* $u(\cdot, 0)$ *has only continuous scattering data, then so has* $u(\cdot, t)$*.*
4. *If the reduced wave function* $m(\cdot, 0, \lambda)$ *has a pole at* $\lambda = \alpha$*, then so has* $m(\cdot, t, \lambda)$ *for all* t*.*
5. $u = [a, m_1]$*, where* m_1 *is the coefficient of* λ^{-1} *in the asymptotic expansion of* $m(\cdot, \cdot, \lambda)$ *at* $\lambda = \infty$*.*

6. Direct scattering for the space-time monopole equation

The linear system associated to the Lax pair (12) for the monopole equation is (14). The first operator $P_1(\mu)$ is a linear operator in spatial variables only. Given a rapidly decaying initial data (A, ϕ) on $\mathbb{R}^2$, the scattering data for the operator $P_1(\mu)$ is the singularity data of the solution E_μ for $P_1(\mu)E_\mu = 0$ satisfying certain boundary condition.

Definition 6.1. A rapidly decaying spatial pair $(A, \phi) : \mathbb{R}^2 \to \bigoplus_{i=1}^4 su(n)$ is said to have only *continuous scattering data* if there exists $E_\mu : \mathbb{R}^2 \to \mathrm{GL}(n,\mathbb{C})$ defined on $\mathcal{O}_\varepsilon^\pm = \{\mu \in \mathbb{C} \mid 1 < |\mu|^{\pm 1} < 1 + \varepsilon\}$ for some $\varepsilon > 0$ such that

1. $\begin{cases} P_1(\mu)E_\mu = \left(\frac{\mu-\mu^{-1}}{2}\nabla_x - \frac{i(\mu+\mu^{-1})}{2}\nabla_y - i\phi\right)E_\mu = 0, \\ E_\mu(\infty) = \mathrm{I}, \quad E_{\bar{u}^{-1}} = (E_\mu^*)^{-1}, \end{cases}$
2. $\mu \mapsto E_\mu(x,t)$ are holomorphic in $\mu \in \mathcal{O}_\varepsilon^\pm$,
3. the limits $\lim_{\mu \in \mathcal{O}^\pm, \mu \to e^{i\theta}} E_\mu = S_\theta^\pm$ exist.

It follows from the reality condition that $S_\theta^- = (S_\theta^+)^{*^{-1}}$. We call the non-negative Hermitian matrix

$$S_\theta = (S_\theta^-)^{-1}S_\theta^+ = (S_\theta^+)^* S_\theta^+$$

the *scattering matrix* or the *continuous scattering data*.

Let $W^{2,1}$ denote the space of maps f whose partial derivatives up to second order are in L^1.

Theorem 6.2 ([32], [16], [13]). *Assume that* (A, ϕ) *is a rapidly decaying spatial data and* (A, ϕ) *is small in* $W^{2,1}$*. Then the continuous scattering matrix* S_θ *exists,* $\mathrm{I} - S_\theta$ *decays for each* θ*, and the scattering matrix* S_θ *satisfies*

(a) $\mathrm{I} - S_\theta$ *is small in* L^∞,

(b) $S_\theta^* = S_\theta \geq 0$,

(c) $\big(-\sin\theta\frac{\partial}{\partial x} + \cos\theta\frac{\partial}{\partial y}\big)S_\theta = 0$.

Theorem 6.3 ([32], [16], [13]). *If* (A, ϕ) *is a smooth solution of the space-time monopole equation in* $\mathbb{R}^2 \times (T_1, T_2)$ *and decays in spatial variables, and has a smooth continuous scattering data. Then*

$$0 = \left(\frac{\partial}{\partial t} + \cos\theta\frac{\partial}{\partial x} + \sin\theta\frac{\partial}{\partial y}\right) S_\theta.$$

Moreover, two gauge equivalent solutions give rise to the same scattering data.

Corollary 6.4. *Let* (A, ϕ) *be as in Theorem* 6.3. *Then there is a unique* $s_0 \colon \mathbb{R} \times S^1 \to \mathrm{GL}(n, \mathbb{C})$ *such that* $s_0^* = s_0$, $s_0(r, e^{i\theta})$ *is rapidly decaying in* $r \in \mathbb{R}$, *and the continuous scattering data for* $(A(\,\cdot\,,\,\cdot\,,t), \phi(\,\cdot\,,\,\cdot\,,t))$ *is*

$$S_\theta(x, y, t) = s_0(x\cos\theta + y\sin\theta - t, e^{i\theta}).$$

7. Inverse scattering for the SU(n)-hierarchy via loop group factorizations

Given $u \in \mathcal{S}(\mathbb{R}, V_n)$, the scattering data for the operator $L_u = d_x + a\lambda + u$ is the singularities of the reduced wave function, which contains two parts, the continuous (jumping line) and the discrete (poles) scattering data. The inverse scattering, which constructs u from the scattering data of L_u, was done in [39], [4].

By Theorem 5.1, the scattering data only depends on $f(\lambda) = m(0, 0, \lambda)$, where m is the reduced wave function. We identify the image of the scattering transform for those u's with only continuous scattering data as a homogeneous space, and then use Pressley–Segal loop group factorization to construct the inverse scattering transform ([27]).

Let $\mathcal{D}_-$ denote the group of smooth $f \colon \mathbb{R} \to \mathrm{GL}(n, \mathbb{C})$ such that

(i) f is the boundary value of a holomorphic map in $\mathbb{C}_+ = \{\lambda \in \mathbb{C} \mid \mathrm{Im}(\lambda) > 0\}$,

(ii) f has the same asymptotic expansion at $r = \pm\infty$,

(iii) decompose $f(r) = p(r)v(r)$ with $p(r)$ upper triangular and $v(r)$ unitary, then $p - \mathrm{I}$ is rapidly decaying.

Suppose u is a solution of the (b, j)-flow (2) in the SU(n)-hierarchy with only continuous scattering data, and $m(x, t, \lambda)$ the reduced wave function for u. Then

$m(x,t,\cdot) \in \mathcal{D}_-$. Set $f(\lambda) = m(0,0,\lambda)$. Since $\psi(x,t,\lambda) = e^{a\lambda x + b\lambda^j t} m(x,t,\lambda)$ satisfies (17),

$$E(x,t,\lambda) = f(\lambda)^{-1} e^{(a\lambda x + a\lambda^2 t)} m(x,t,\lambda), \tag{18}$$

is a solution of (17) with $E(0,0,\lambda) = \mathrm{I}$. Because the right hand side of (18) is holomorphic in $\mathbb{C}_+$, $E(x,t,\lambda)$ is holomorphic in $\lambda \in \mathbb{C}_+$. Proposition 2.2 implies that E satisfies the $U(n)$-reality condition $E(x,t,\bar{\lambda})^* E(x,t,\lambda) = \mathrm{I}$. So by the reflection principal $E(x,t,\lambda)$ is holomorphic for all $\lambda \in \mathbb{C}$.

Set $e_{a,1}(x)(\lambda) = e^{a\lambda x}$, $e_{b,j}(t) = e^{b\lambda^j t}$, $E(x,t)(\lambda) = E(x,t,\lambda)$, and $m(x,t)(\lambda) = m(x,t,\lambda)$. Then we can rewrite (18) as

$$f^{-1} e_{a,1}(x) e_{b,j}(t) = E(x,t) m(x,t)^{-1}. \tag{19}$$

Here $f, m(x,t) \in \mathcal{D}_-$, and $e_{a,1}(x)e_{b,j}(t)$ and $E(x,t)$ holomorphic in $\mathbb{C}$ and satisfy the $U(n)$-reality condition. To construct the inverse scattering is to construct m from f. In other words, given $f \in \mathcal{D}_-$, we want to find a method to factor $f^{-1} e_{a,1}(x) e_{b,j}(t)$ as $E(x,t)m^{-1}(x,t)$ such that $E(x,t)$ satisfies the $U(n)$-reality condition and is holomorphic in $\mathbb{C}$ and $m(x,t) \in \mathcal{D}_-$ for all (x,t). We need Pressley–Segal loop group factorization [22] given below to do this factorization.

Let $S^2 \setminus S^1 = \mathbb{C} \cup \{\infty\} = \Omega_+ \cup \Omega_-$, where $\Omega_+ = \{\mu \in \mathbb{C} \mid |\mu| < 1\}$ and $\Omega_- = \{\lambda \in S^2 \mid |\mu| > 1\}$. Let $\Lambda(\mathrm{SL}(n,\mathbb{C}))$ denote the group of smooth loops $g\colon S^1 \to \mathrm{SL}(n,\mathbb{C})$, and $\Lambda_+(\mathrm{SL}(n,\mathbb{C}))$ the subgroup of $g \in \Lambda(\mathrm{SL}(n,\mathbb{C}))$ such that g can be extended to a holomorphic map on Ω_+ and $g(-1)$ is upper triangular with real diagonal entries. Let $\Lambda(\mathrm{SU}(n))$ denote the loops in $\mathrm{SU}(n)$. The Pressley–Segal factorization is the analogue of the Iwasawa decomposition of $\mathrm{SL}(n,\mathbb{C})$ for loop groups:

Theorem 7.1 (Pressley–Segal Factorization Theorem [22]). *The multiplication map from $\Lambda(\mathrm{SU}(n)) \times \Lambda_+(\mathrm{SL}(n,\mathbb{C}))$ to $\Lambda(\mathrm{SL}(n,\mathbb{C}))$ is a bijection. In particular, given $f \in \Lambda(\mathrm{SL}(n,\mathbb{C}))$, there exist unique $g \in \Lambda(\mathrm{SU}(n))$ and $h_+ \in \Lambda_+(\mathrm{SL}(n,\mathbb{C}))$ such that $f = gh_+$.*

If we change the spectral parameter λ by the linear fractional transformation $\mu = \frac{1+i\lambda}{1-i\lambda}$, then we can see that $\mathcal{D}_-$ is isomorphic to a subgroup of $\Lambda_+(\mathrm{SL}(n,\mathbb{C}))$. In fact, we have

Proposition 7.2 ([27]). *Given a map $g\colon S^1 \to \mathrm{GL}(n,\mathbb{C})$, let $\Phi(g)\colon \mathbb{R} \to \mathrm{GL}(n,\mathbb{C})$ be the map defined by $\Phi(g)(r) = g\left(\frac{1+ir}{1-ir}\right)$. Then:*

1. *g is smooth if and only if $\Phi(g)$ is smooth and has the same asymptotic expansion at $r = \pm\infty$.*

2. *$j_\infty(g - \mathrm{I})_{-1} = 0$ (the infinite jet of $g - \mathrm{I}$ at $\mu = -1$) if and only if $\Phi(g) - \mathrm{I}$ is rapidly decaying.*

3. *Suppose g extends holomorphically to $|\mu| < 1$, and define g on $|\mu| > 1$ by $g(\mu) = (g(\bar{\mu}^{-1})^*)^{-1}$. Then $f(\lambda) = g\left(\frac{1+i\lambda}{1-i\lambda}\right)$ is holomorphic in $\lambda \in \mathbb{C} \setminus \mathbb{R}$ and satisfies the reality condition $f(\bar{\lambda})^* f(\lambda) = \mathrm{I}$.*

Corollary 7.3. *$\mathcal{D}_-$ is isomorphic to the subgroup of $g \in \Lambda_+(\mathrm{SL}(n,\mathbb{C}))$ such that $j_\infty(h - \mathrm{I})_{-1} = 0$ where $g = hv$ with h upper triangular and v unitary.*

Now we go back to the problem of factorizing $f^{-1}e_{a,1}(x)$. By Proposition 7.2, $\Phi^{-1}(f^{-1}e_{a,1}(x))$ does not belong to $\Lambda(\mathrm{SL}(n,\mathbb{C}))$. So we can not use Theorem 7.1 to do the factorization directly. However, if we write $f = pv$ with p upper triangular and v unitary, then by definition of $\mathcal{D}_-$, $p - \mathrm{I}$ is rapidly decaying. This implies that $\Phi^{-1}(e_{a,1}^{-1}(x)p^{-1}e_{a,1}(x))$ lies in $\Lambda(\mathrm{SL}(n,\mathbb{C}))$. Apply the Pressley–Segal loop group factorization to get

$$e_{a,1}^{-1}(x)p^{-1}e_{a,1}(x) = B(x)m(x)^{-1}$$

such that $\Phi^{-1}(B(x)) \in \Lambda(\mathrm{SU}(n))$ and $\Phi^{-1}(m(x)) \in \Lambda_+(\mathrm{SL}(n,\mathbb{C}))$. Since $p(\lambda)$, $e_{a,1}(x)(\lambda)$, and $m(x)(\lambda)$ are smooth for $\lambda \in \mathbb{R}$ and can be extended holomorphically to $\lambda \in \mathbb{C}_+$, so is $B(x)(\lambda)$. But $B(x)(\lambda)$ is unitary for $\lambda \in \mathbb{R}$ implies that $B(x)(\lambda)$ can be extended holomorphically across the real axis in the λ-plane by defining $B(x)(\lambda) = (B(x)(\bar{\lambda})^*)^{-1}$. Hence $\lambda \mapsto B(x)(\lambda)$ is holomorphic for all $\lambda \in \mathbb{C}$. Therefore

$$\begin{aligned} f^{-1}e_{a,1}(x) = v^{-1}p^{-1}e_{a,1}(x) &= v^{-1}e_{a,1}(x)(e_{a,1}^{-1}(x)p^{-1}e_{a,1}(x)) \\ &= v^{-1}e_{a,1}(x)B(x)m(x)^{-1} = E(x)m(x)^{-1}. \end{aligned}$$

But $E(x)(\lambda) = v^{-1}(\lambda)e^{a\lambda x}B(x)(\lambda)$ is holomorphic for $\lambda \in \mathbb{C}$.

Since $E(x,\lambda) = f^{-1}(\lambda)e^{a\lambda x}m(x,\lambda)$,

$$E^{-1}\partial_x E = m^{-1}\partial_x m + m^{-1}a\lambda m.$$

Use the asymptotic expansion at $\lambda = \infty$ to conclude that $E^{-1}\partial_x E$ must be a degree one polynomial in λ. So if $m_1(x)$ is the coefficient of λ^{-1} in the asymptotic expansion of $m(x,\lambda)$ at $\lambda = \infty$, then

$$E^{-1}\partial_x E = a\lambda + u_f, \quad \text{where } u_f = [a, m_1].$$

Note that the scattering data of $d_x + a\lambda + u_f$ is $e^{-a\lambda x}f_+f_-^{-1}e^{a\lambda x}$. However, the map $\mathcal{F}(f) = u_f$ is not one to one. In fact, $u_{f_1} = u_{f_2}$ if and only if there is $h \in \mathcal{D}_-$ such that $h(r)$ is diagonal for all $r \in \mathbb{R}$. These give a rough idea of how the following results are obtained.

Theorem 7.4 ([27]). *Assume a, b are diagonal matrices in $su(n)$, and a has distinct eigenvalues. If $f \in \mathcal{D}_-$, then there exist $E(x,t,\lambda)$ and $m(x,t,\lambda)$ such that*

1. $f^{-1}e_{a,1}(x)e_{b,j}(t) = E(x,t,\cdot)m(x,t,\cdot)^{-1}$,
2. *E is holomorphic for $\lambda \in \mathbb{C}$, $E(x,t,\bar{\lambda})^*E(x,t,\lambda) = \mathrm{I}$, and $m(x,t,\cdot) \in \mathcal{D}_-$,*

3. *$u_f = [a, m_1]$ is a solution of the (b, j)-flow equation* (2) *in the* $\mathrm{SU}(n)$*-hierarchy, and E is the frame for the Lax pair associated to u with initial condition $E(0, \lambda) = \mathrm{I}$, where $m_1(x, t)$ is the coefficient of λ^{-1} in the asymptotic expansion of $m(x, t, \lambda)$ at $\lambda = \infty$,*

4. *$u_f(x, t)$ is defined for all $(x, t) \in \mathbb{R}^2$ and is rapidly decaying in x for each t,*

5. *if f also satisfies the* $\mathrm{SU}(n)/\mathrm{SO}(n)$*-reality condition, then u_f is a solution of the (b, j)-flow in the* $\mathrm{SU}(n)/\mathrm{SO}(n)$*-hierarchy.*

Theorem 7.5 ([27]). *Let $\mathcal{S}^c(\mathbb{R}, V_n)$ denote the space of all $u \in \mathcal{S}(\mathbb{R}, V_n)$ such that $L_u = d_x + a\lambda + u$ has only continuous scattering data, and $\mathcal{D}_-(A)$ denotes the subgroup of $f \in \mathcal{D}_-$ such that $f(r)$ is diagonal for all $r \in \mathbb{R}$, and $\mathcal{F} : \mathcal{S}^c(\mathbb{R}, V_n) \to \mathcal{D}_-/\mathcal{D}_-(A)$ defined by $\mathcal{F}(u) = [m(0, \cdot)]$, where $m(x, \lambda)$ is the reduced wave function of $L_u = d_x + a\lambda + u$. Then $\mathcal{F}$ is a bijection, and $\mathcal{F}^{-1}([f]) = [a, m_1]$, where m_1 is the coefficient of λ^{-1} in the asymptotic expansion of m at $\lambda = \infty$.*

Theorem 7.6 ([27]). *Let $L^\tau_+(\mathrm{SL}(n, \mathbb{C}))$ denote the group of holomorphic maps $f : \mathbb{C} \to \mathrm{GL}(n, \mathbb{C})$ that satisfy the* $\mathrm{SU}(n)$*-reality condition, and $\mathcal{D}_+(A)$ the subgroup of $L^\tau_+(\mathrm{SL}(n, \mathbb{C}))$ generated by $\{e_{b,j}(t) \mid b \in su(n)$ diagonal, $j \geq 1$ integer$\}$. Then $\mathcal{D}_+(A)$ acts on $\mathcal{D}_-/\mathcal{D}_-(A)$ by $e_{a,j}(t) * [f] = [m(t)]$, where $m(t)$ is obtained by factoring $f^{-1}e_{b,j}(t) = E(t)m(t)^{-1}$ such that $E(t) \in L^\tau_+(\mathrm{SL}(n, \mathbb{C}))$ and $m(t) \in \mathcal{D}_-$. Moreover, the (b, j)-flow in the* $\mathrm{SU}(n)$*-hierarchy corresponds to the action of $e_{b,j}(t)$ on $\mathcal{D}_-/\mathcal{D}_-(A)$ under the isomorphism $\mathcal{F}$.*

Theorem 7.7 ([27]). *Let $a_1, a_2, \ldots, a_n$ be linearly independent diagonal matrices in $u(n)$, and $f \in \mathcal{D}_-$. Then we can factor*

$$f^{-1}e_{a_1,1}(x_1) \ldots e_{a_n,1}(x_n) = E(x)m(x)^{-1}$$

such that $E(x) \in L^\tau_+(\mathrm{GL}(n, \mathbb{C}))$ and $m(x) \in \mathcal{D}_-$. Moreover,

1. *$v = m_1^\perp$ is a solution of the $U(n)$-system, where $m_1(x)$ is the coefficient of λ^{-1} in the asymptotic expansion of $m(x)(\lambda)$ at $\lambda = \infty$ and $\xi^\perp = \xi - \sum_{i=1}^n \xi_{ii} e_{ii}$,*

2. *if $f \in \mathcal{D}_-$ satisfies the $\frac{U(n)}{O(n)}$-reality condition, then $v = (m_1)^\perp$ is a solution of the $\frac{U(n)}{O(n)}$-system.*

In other words, the $U(n)$-system is the system obtained by putting together the $(a_1, 1)$-, …, $(a_n, 1)$-flow in the $U(n)$-hierarchy.

8. The inverse scattering for monopole equations

The scattering data of the linear operator

$$P_1(\mu) = \mu\nabla_z - \mu^{-1}\nabla_{\bar{z}} - i\phi$$

on the $z = x + iy$ plane for the rapidly decaying spatial pair (A, ϕ) is a smooth map $s_0 : \mathbb{R}^1 \times S^1 \to \mathrm{GL}(n, \mathbb{C})$. The inverse scattering for $P_1(\mu)$, which constructs $(A, \phi) : \mathbb{R}^2 \to \oplus^4 su(n)$ from s_0, was done in [32], [16]. In this section, we give a brief review of the construction of the inverse scattering transform for $P_1(\mu)$ via Pressley–Segal loop group factorization given in [13].

Theorem 8.1 ([13]). *Suppose $s : \mathbb{R} \times S^1 \to \mathrm{GL}(n, \mathbb{C})$ is smooth such that $s^* = s \geq 0$ and $s(r, e^{i\theta}) - \mathrm{I}$ is rapidly decaying for $r \in \mathbb{R}$. Define*

$$S(x, y, t, e^{i\theta}) = s(x \cos\theta + y \sin\theta - t, e^{i\theta}).$$

Then there exists a smooth $E : \mathbb{R}^{2,1} \times (\mathbb{C} \setminus S^1) \to \mathrm{GL}(n, \mathbb{C})$ such that

1. *$E^*_{\bar{\mu}^{-1}} E_\mu = \mathrm{I}$, where $E_\mu = E(\ldots, \mu)$,*
2. *$((\partial_t + \mu\partial_z)E_\mu)E_\mu^{-1} = B_0 + \mu B_1$ and $((\partial_t + \mu^{-1}\partial_{\bar{z}})E_\mu)E_\mu^{-1} = -(B_0^* + \mu^{-1}B_1^*)$ for some $B_0, B_1 : \mathbb{R}^{2,1} \to sl(n, \mathbb{C})$,*
3. *if we set $A_z = \frac{1}{2}B_1$, $A_t = \frac{1}{2}(B_0 - B_0^*)$, $\phi = \frac{i}{2}(B_0 + B_0^*)$, then (A, ϕ) is a solution of the space-time monopole equation decaying rapidly in the spatial variables,*
4. *the scattering data for $(A(\cdot, \cdot, t), \phi(\cdot, \cdot, t))$ is $S(\cdot, \cdot, t, \cdot)$.*

Here is a sketch of the proof: Set $S_\mu = S(\ldots, \mu)$ for $|\mu| = 1$. Write $S_\mu = P_\mu^2$ with $P_\mu^* = P_\mu$. By Pressley–Segal Factorization Theorem 7.1 we can factor $P_\mu = U_\mu E_\mu^+$ with $U_\cdot$ a loop in $\mathrm{SU}(n)$ and E_μ^+ extends holomorphically to $|\mu| < 1$. Define $E_\mu^- = ((E^+_{\bar{\mu}^{-1}})^*)^{-1}$ for $|\mu| > 1$. Then $S_\mu = (E_\mu^-)^{-1}E_\mu^+$. The rest of the theorem can be proved using the fact that $(-\sin\theta\ \partial_x + \cos\theta\ \partial_y)S = 0$ and $(\cos\theta\ \partial_x + \sin\theta\ \partial_y - \partial_t)S = 0$.

Corollary 8.2 ([13]). *Suppose (A_0, ϕ_0) is a rapidly decaying spatial pair with only continuous scattering data. Then there is a global solution (A, ϕ) of the space-time monopole equation decaying rapidly in the spatial variables such that the scattering data of $(A(\cdot, \cdot, 0), \phi(\cdot, \cdot, 0))$ and (A_0, ϕ_0) are the same. Moreover, any two such solutions are gauge equivalent.*

Corollary 8.3. *There is a bijective correspondence between the space of solutions of the space-time monopole equation with only continuous scattering data modulo the gauge group, and the group of maps $f : \mathbb{R} \to \Lambda^\tau_+(\mathrm{SL}(n, \mathbb{C}))$ such that $(f^* f)(r)(e^{i\theta}) - \mathrm{I}$ is rapidly decaying in $r \in \mathbb{R}$.*

9. Birkhoff factorization and local solutions

The factorization (19)

$$f^{-1}(\lambda)e^{a\lambda x + b\lambda^j t} = E(x, t, \lambda)m(x, t, \lambda)^{-1}$$

is for $f, m(x,t,\cdot)$ in $\mathcal{D}_-$ and E holomorphic in $\lambda \in \mathbb{C}$ and satisfies the $U(n)$-reality condition. However if f is holomorphic at $\lambda = \infty$, then we can use the Birkhoff factorization to get E and m such that E is holomorphic in $\mathbb{C}$ and m is holomorphic at $\lambda = \infty$. Moreover, it can be shown easily that E is the frame of a solution of the (b, j)-flow with $E(0,0,\lambda) = \mathrm{I}$. Since the Birkhoff factorization only works on an open dense subset of loops, solutions constructed this way are local solutions defined in a neighborhood of $(0,0)$.

Let $\varepsilon > 0$, and $\mathcal{O}_\varepsilon = \{\lambda \mid |\lambda| > \frac{1}{\varepsilon}\}$ an open neighborhood of ∞ in $S^2 = \mathbb{C} \cup \{\infty\}$. Then $S^2 = C \cup \mathcal{O}_\varepsilon$. Let $L^\tau(\mathrm{SL}(n,\mathbb{C}))$ denote the group of holomorphic maps f from $C \cap \mathcal{O}_\infty$ to $\mathrm{SL}(n,\mathbb{C})$ satisfying the $\mathrm{SU}(n)$-reality condition $f(\bar{\lambda})^* f(\lambda) = \mathrm{I}$, $L^\tau_+(\mathrm{SL}(n,\mathbb{C}))$ the subgroup of $f \in L^\tau(\mathrm{SL}(n,\mathbb{C}))$ that extend holomorphically to $\mathbb{C}$, and $L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$ the subgroup of $f \in L^\tau(\mathrm{SL}(n,\mathbb{C}))$ that extend holomorphically to $\mathcal{O}_\varepsilon$ and satisfying $f(\infty) = I$.

Theorem 9.1 (Birkhoff Factorization Theorem (cf. [22])). *The multiplication map $L^\tau_+(\mathrm{SL}(n,\mathbb{C})) \times L^\tau_-(\mathrm{SL}(n,\mathbb{C})) \to L^\tau(\mathrm{SL}(n,\mathbb{C}))$ is injective and the image is an open dense subset of $L^\tau(\mathrm{SL}(n,\mathbb{C}))$.*

Let a, b be diagonal matrices in $su(n)$ such that a is regular. Then $e_{a,1}(x)e_{b,j}(t) \in L^\tau_+(\mathrm{SL}(n,\mathbb{C}))$. Given $f \in L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$, by the Birkhoff factorization there exists $\delta > 0$ such that

$$f^{-1}e_{a,1}(x)e_{b,j}(t) = E(x,t)m(x,t)^{-1}$$

with $E(x,t) \in L^\tau_+(\mathrm{SL}(n,\mathbb{C}))$ and $m(x,t) \in L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$ for all $(x,t) \in B_\delta(0)$. Here $B_\delta(0)$ is the ball of radius δ centered at $(0,0)$. Then

$$\begin{cases} E^{-1}\partial_x E = m^{-1}m_x + m^{-1}a\lambda m, \\ E^{-1}\partial_t E = m^{-1}m_x + m^{-1}b\lambda^j m. \end{cases}$$

Since m is holomorphic at $\lambda = \infty$ and $m(x,t)(\infty) = \mathrm{I}$, $E^{-1}\partial_x E$ and $E^{-1}\partial_t E$ must be a polynomial of degree 1 and j in λ respectively. Hence E must be a frame of a solution of the (b, j)-flow in the $\mathrm{SU}(n)$-hierarchy. So we have

Theorem 9.2 ([27]). *If $f \in L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$, then there exist an open neighborhood $\mathcal{O}$ of $(0,0)$, $E(x,t) \in L^\tau_+(\mathrm{SL}(n,\mathbb{C}))$, and $m(x,t) \in L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$ such that $f^{-1}e_{a,1}(x)e_{b,j}(t) = E(x,t)m(x,t)^{-1}$ for all $(x,t) \in \mathcal{O}$. Moreover,*

1. *$u = [a, m_1]$ is a solution of the (b, j)-flow in the $\mathrm{SU}(n)$-hierarchy, where $m_1(x,t)$ is the coefficient of λ^{-1} the expansion of $m(x,t)(\lambda)$ at $\lambda = \infty$, (we will use $f * 0$ to denote u),*
2. *E is the frame of the Lax pair of u such that $E(0)(\lambda) = \mathrm{I}$,*
3. *if f satisfies the $\mathrm{SU}(n)/\mathrm{SO}(n)$-reality condition, then $u = [a, m_1]$ is a solution of the (b, j)-flow in the $\mathrm{SU}(n)/\mathrm{SO}(n)$-hierarchy.*

Note that for $f \in \mathcal{D}_-$, Theorem 7.4 gives a global solution u_f of the (b, j)-flow on $\mathcal{S}(\mathbb{R}, V_n)$ in the $\mathrm{SU}(n)$-hierarchy. But for $f \in L^\tau_-(\mathrm{SL}(n,\mathbb{C}))$, the above theorem only gives a local solution of the (b, j)-flow in general.

Theorem 9.3 ([27]). *If $f \in L^\tau_-(\mathrm{GL}(n,\mathbb{C}))$, then there exist an open neighborhood $\mathcal{O}$ of 0 in $\mathbb{R}^n$, $E(x) \in L^\tau_+(\mathrm{GL}(n,\mathbb{C}))$ and $m(x) \in L^\tau_-(\mathrm{GL}(n,\mathbb{C}))$ for $x \in \mathcal{O}$ such that $f^{-1}e_{a_1,1}(x_1)\dots e_{a_n,1}(x_n) = E(x)m(x)^{-1}$. Moreover,*

1. *$v = m_1^\perp$ is a solution of the $U(n)$-system, where $m_1(x)$ is the coefficient of λ^{-1} in the expansion of $m(x)(\lambda)$ at $\lambda = \infty$ and $m_1^\perp = m_1 - \sum_i (m_1)_{ii}e_{ii}$, (we will use $f * 0$ to denote v),*
2. *E is the frame of the Lax pair (6) of v such that $E(0)(\lambda) = \mathrm{I}$,*
3. *If f also satisfies the $U(n)/O(n)$-reality condition, then $v = f*0$ is a solution of the $U(n)/O(n)$-system.*

10. Bäcklund transformations for the $U(n)$-hierarchy

In general, the solution $f * 0$ constructed in Theorem 9.2 has singularities. But if f is rational, then $f * 0$ is a global solution of the (b, j)-flow in the $\mathrm{SU}(n)$-hierarchy, and can be computed explicitly. These are the soliton solutions. Moreover, if f is rational with only one simple pole and E is a frame of a solution u, then the Birkhoff factorization $fE = \tilde{E}\tilde{f}$ can be carried out by an explicit algebraic algorithm so that $\tilde{E}$ is a frame of the new solution. These give Bäcklund transformations for the (b, j)-flow.

If $\alpha \in \mathbb{C} \setminus \mathbb{R}$ and π is a Hermitian projection of $\mathbb{C}^n$, then the map

$$g_{\alpha,\pi}(\lambda) = \mathrm{I} + \frac{\alpha - \overline{\alpha}}{\lambda - \alpha}\,\pi$$

satisfies the $U(n)$-reality condition. So $g_{\alpha,\pi} \in L^\tau_-(\mathrm{GL}(n,\mathbb{C}))$.

The following theorem is a key ingredient for constructing Bäcklund transformations for the (b, j)-flow in the $\mathrm{SU}(n)$-hierarchy.

Theorem 10.1 ([29]). *Given $f \in L^\tau_+(\mathrm{SL}(n,\mathbb{C}))$ and $g_{\alpha,\pi}$, let $\tilde{\pi}$ be the Hermitian projection of $\mathbb{C}^n$ onto $f(\alpha)^{-1}(\mathrm{Im}\,\pi)$. Then $g_{\alpha,\pi}f = \tilde{f}g_{\alpha,\tilde{\pi}}$ and $\tilde{f} \in L^\tau_+(\mathrm{SL}(n,\mathbb{C})$.*

Proof. Set $\tilde{f}(\lambda) = g_{\alpha,\pi}(\lambda)f(\lambda)g_{\alpha,\tilde{\pi}}(\lambda)^{-1} = (\mathrm{I} + \frac{\alpha-\overline{\alpha}}{\lambda-\alpha}\pi^\perp)f(\lambda)(\mathrm{I} + \frac{\overline{\alpha}-\alpha}{\lambda-\overline{\alpha}}\tilde{\pi}^\perp)$. Note that $\tilde{f}$ is holomorphic for $\lambda \in \mathbb{C} \setminus \{\alpha, \overline{\alpha}\}$. But

$$\mathrm{Res}(\tilde{f}, \alpha) = (\alpha - \overline{\alpha})\pi^\perp f(\alpha)\tilde{\pi}, \quad \mathrm{Res}(\tilde{f}, \overline{\alpha}) = (\overline{\alpha} - \alpha)\pi f(\overline{\alpha})\tilde{\pi}^\perp.$$

By definition $f(\alpha)(\mathrm{Im}\,\tilde{\pi}) = \mathrm{Im}\,\pi$, so $\tilde{f}$ is holomorphic at $\lambda = \alpha$. Set $V = \mathrm{Im}\,\pi$ and $\tilde{V} = \mathrm{Im}\,\tilde{\pi}$. Since f satisfies the reality condition, we have

$$(f(\overline{\alpha})(\tilde{V}^\perp), V) = (\tilde{V}^\perp, f(\overline{\alpha})^*(V)) = (\tilde{V}^\perp, f(\alpha)^{-1}(V)) = (\tilde{V}^\perp, \tilde{V}) = 0.$$

This implies that $\operatorname{Res}(\tilde{f}, \overline{\alpha}) = 0$, hence $\tilde{f}$ is holomorphic in $\mathbb{C}$. □

The proof of the above theorem in fact gives the following more general result:

Theorem 10.2 ([12]). *Let $\mathcal{O}$ be an open subset of $\mathbb{C}$ that is invariant under complex conjugation, and $f: \mathcal{O} \to \mathrm{SL}(n, \mathbb{C})$ a meromorphic map satisfying the $U(n)$-reality condition. Let $\alpha \in \mathbb{C} \setminus \mathbb{R}$, and π a Hermitian projection of $\mathbb{C}^n$. Suppose f is holomorphic and non-singular at $\lambda = \alpha$. Let $\tilde{\pi}$ denote the Hermitian projection of $\mathbb{C}^n$ onto $f(\alpha)^{-1}(\operatorname{Im}\pi)$. Then $g_{\alpha,\pi} f = \tilde{f} g_{\alpha,\tilde{\pi}}$, and $\tilde{f}$ is holomorphic and non-degenerate at $\lambda = \alpha$ and satisfies the $U(n)$-reality condition.*

Theorem 10.3 (Bäcklund transformation for the (b, j)-flow, [29]). *Suppose u is a solution of the (b, j)-flow* (2) *in the* $\mathrm{SU}(n)$*-hierarchy, and $E(x, t, \lambda)$ is the frame for the Lax pair of u such that E is holomorphic for $\lambda \in \mathbb{C}$ and $E(0, 0, \lambda) = \mathrm{I}$. Given $\alpha \in \mathbb{C} \setminus \mathbb{R}$ and a Hermitian projection π of $\mathbb{C}^n$, set $\tilde{\pi}(x, t)$ to be the Hermitian projection of $E(x, t, \alpha)^{-1}(\operatorname{Im}\pi)$, $\tilde{E} = g_{\alpha,\pi} E g_{\alpha,\tilde{\pi}}^{-1}$, and $\tilde{u} = u + (\alpha - \overline{\alpha})[a, \tilde{\pi}]$. Then*

1. *$\tilde{u}$ is again a solution of* (2)*, and $\tilde{E}$ is the frame of $\tilde{u}$,*
2. *if u is smooth for all $(x, t) \in \mathbb{R}^2$, then so is $\tilde{u}$,*
3. *if $u(x, t)$ is rapidly decaying in x for all t, then so is $g_{\alpha,\pi} * u$.*

Let $\mathcal{D}_-^r$ denote the group of rational maps $g: S^2 \to \mathrm{GL}(n, \mathbb{C})$ that satisfy the $U(n)$-reality condition and $g(\infty) = \mathrm{I}$. Uhlenbeck proved in [30] that $\mathcal{D}_-^r$ is generated by the set $\{g_{\alpha,\pi} \mid \alpha \in \mathbb{C} \setminus \mathbb{R},\ \pi \text{ is a Hermitian projection of } \mathbb{C}^n\}$.

A *pure soliton* for the (b, j)-flow is a solution that is rapidly decaying in the spatial variable, has no continuous scattering data and has finitely many discrete scattering data, so its reduced wave function $m(x, \lambda)$ is rational in λ. Or equivalently, its reduced wave function m lies in the group $\mathcal{D}_-^r$.

Corollary 10.4 ([29]). *The group $\mathcal{D}_-^r$ acts on the space of solutions of the (b, j)-flow in the* $\mathrm{SU}(n)$*-hierarchy. In fact, if $g = g_{\alpha_1,\pi_1} \dots g_{\alpha_k,\pi_k}$, then $g * u = g_{\alpha_1,\pi_1} * (\cdots * (g_{\alpha_k,\pi_k} * u) \cdots)$.*

Corollary 10.5 ([27]). *Let $E_0(x, t, \lambda) = e^{a\lambda x + b\lambda^j t}$ (E_0 is the frame for the vacuum solution $u = 0$ of the (b, j)-flow). If we apply BT (Theorem* 10.3*) to E_0 repeatedly, then we obtain all pure soliton solutions of the (b, j)-flow, i.e., solutions with continuous scattering data $S = \mathrm{I}$ and finitely many discrete scattering data.*

Theorem 10.6 ([27]). *Let u be the global solution of the (b, j)-flow* (2) *in the* $\mathrm{SU}(n)$*-hierarchy constructed in Theorem* 7.4 *with only continuous scattering data, and E its frame. If we apply BT (Theorem* 10.3*) repeatedly to E, then we obtain solutions of the (b, j)-flow that have both continuous and finite discrete scattering data. Conversely, any solution u of the (b, j)-flow in the* $\mathrm{SU}(n)$*-hierarchy that has continuous scattering data and finite discrete scattering data can be constructed this way.*

Theorem 10.7 (Bäcklund transformation for the $U(n)$-system, [29]). *Suppose v is a solution of the $U(n)$-system, and $E(x,\lambda)$ is the frame for the Lax pair of v such that E is holomorphic for $\lambda \in \mathbb{C}$ and $E(0,\lambda) = \mathrm{I}$. Given $\alpha \in \mathbb{C} \setminus \mathbb{R}$ and a Hermitian projection π of $\mathbb{C}^n$, set $\tilde{\pi}(x)$ to be the Hermitian projection of $E(x,\alpha)^{-1}(\operatorname{Im}\pi)$, $\tilde{E} = g_{\alpha,\pi} E g_{\alpha,\tilde{\pi}}^{-1}$, and $\tilde{v} = v + (\alpha - \overline{\alpha})\tilde{\pi}^{\perp}$, where $\xi^{\perp} = \xi - \sum_{i=1}^{n} \xi_{ii} e_{ii}$. Then*

1. *$\tilde{v}$ is again a solution of the $U(n)$-system, and $\tilde{E}$ is a frame of the Lax pair of $\tilde{v}$,*
2. *if v is smooth for all $x \in \mathbb{R}^n$, then so is $\tilde{v}$.*

The group $\mathcal{D}_-^r$ acts on the space of solutions of the $U(n)$-system such that $g_{\alpha,\pi} * v = \tilde{v}$, where $\tilde{v}$ is given in the above theorem.

It is easy to see that if u is a solution of the $\frac{U(n)}{O(n)}$-system, $s \in \mathbb{R}$, and $\overline{\pi} = \pi$, then $g_{is,\pi}$ satisfies the $U(n)/O(n)$-reality condition and $g_{is,\pi} * u$ is also a solution of the $\frac{U(n)}{O(n)}$-system. In general,

Corollary 10.8 ([29]). *If $g \in \mathcal{D}_-^r$ satisfies the $\frac{U(n)}{O(n)}$-reality condition and v is a solution of the $\frac{U(n)}{O(n)}$-system, then $g * v$ is again a solution of the $\frac{U(n)}{O(n)}$-system.*

11. Bäcklund transformations for the space-time monopole equation

We use Lax pair (13) to construct soliton solutions and Bäcklund transformations for the monopole equation. Since the spectral parameter λ in (13) is related to the spectral parameter μ in (12) by $\mu = \frac{1-i\lambda}{1+i\lambda}$, the continuous scattering data for (13) is the jump across the real axis, and the discrete scattering data is given by the poles in $\mathbb{C} \setminus \mathbb{R}$ and their residues.

Definition 11.1. A monopole (A, ϕ) rapidly decaying in the spatial variable is a *k-soliton* if there is a gauge equivalent monopole with a frame (solution of (15)) ψ_λ that is rational in λ with k poles, $\psi_\infty = \mathrm{I}$, and $\lim_{\|(x,y)\|\to\infty} \psi_\lambda(x, y, t, \lambda) = h(\lambda)$ is independent of t.

We identify the set of all rank k Hermitian projections of $\mathbb{C}^n$ as the complex Grassmannian $\mathrm{Gr}(k, \mathbb{C}^n)$ by $\pi \mapsto \operatorname{Im}\pi$.

Theorem 11.2 ([33]). *Let $\alpha \in \mathbb{C} \setminus \mathbb{R}$, $\pi_0 \colon S^2 \to \mathrm{Gr}(k, \mathbb{C}^n)$ a holomorphic map, $\xi = \frac{1}{2}(t + x)$, $\eta = \frac{1}{2}(t - x)$, $\pi(x, y, t) = \pi_0(y + \alpha\xi + \alpha^{-1}\eta)$, and*

$$g_{\alpha,\pi}(x, y, t) = \mathrm{I} + \frac{\alpha - \overline{\alpha}}{\lambda - \alpha}\, \pi(x, y, t).$$

Then $g_{\alpha,\pi}$ is a 1-soliton monopole frame. Moreover, all 1-soliton frames are of this form up to gauge equivalence.

The following theorem is a consequence of Theorem 10.2 and the fact that $\lambda\partial_y - \partial_\xi$ and $\lambda\partial_\eta - \partial_y$ are derivations.

Theorem 11.3 (BT for monopoles, [12]). *Suppose $\alpha \in \mathbb{C} \setminus \mathbb{R}$ is a constant, and ψ is a frame of the monopole solution (A, ϕ) (i.e., solution of* (15)*), and $\psi(x, y, t, \tau)$ is holomorphic and non-degenerate at $\tau = \alpha$. Let $g_{\alpha,\pi}$ be a 1-soliton monopole frame, $\tilde{\pi}(x, y, t)$ the Hermitian projection of $\mathbb{C}^n$ onto $\psi(x, y, t, \alpha)(\operatorname{Im}\pi(x, y, t))$, and $\tilde{\psi} = g_{\alpha,\tilde{\pi}}\psi g_{\alpha,\pi}^{-1}$. Then*

1. *$\tilde{\psi}$ is holomorphic and non-degenerate at $\tau = \alpha$,*
2. *$\psi_1 = g_{\alpha,\tilde{\pi}}\psi = \tilde{\psi} g_{\alpha,\pi}$ is a frame for* (15) *with $\tilde{A}$, $\tilde{\phi}$ given by*

$$\begin{cases} \tilde{A}_\eta = A_\eta, \\ \tilde{A}_\xi = (1 - \frac{\bar{a}}{\alpha})(\partial_\xi\tilde{\pi})h + h^{-1}A_\xi h, \\ \tilde{A}_y + \tilde{\phi} = A_y + \phi, \\ \tilde{A}_y - \tilde{\phi} = (1 - \frac{\bar{\alpha}}{\alpha})(\partial_y\tilde{\pi})h + h^{-1}(A_y - \phi)h, \end{cases}$$

   where $h = \tilde{\pi} + \frac{\alpha}{\bar{\alpha}}\,\tilde{\pi}^\perp$,

3. *$(\tilde{A}, \tilde{\phi})$ is a solution of the space-time monopole equation.*

If we apply Theorem 11.3 to a 1-soliton k-times, then we get a $(k + 1)$-soliton whose frame has $(k + 1)$ distinct simple poles. Moreover, we have

Corollary 11.4. *Suppose (A, ϕ) is a solution of the space-time monopole equation with only continuous scattering data and E is its frame constructed in Theorem* 8.1. *If we apply Theorem* 11.3 *to E repeatedly, then we obtain a monopole whose frame has both continuous scattering data and finitely many distinct simple poles.*

Note that a BT for flows in the SU(n)-hierarchy adds to a given solution, a soliton with scattering pole at α, regardless of whether the given solution already has a scattering pole at α or not. But this is not the case for the monopole equation, so BTs produce soliton monopole frames with distinct simple poles only. Ward and his group ([35], [19], [3], [20]) take limits of soliton monopole frames with 2 and 3 distinct poles to construct 2- and 3-soliton monopoles with a double and a triple pole at i that are time dependent. Dai and Terng used BT (Theorem 11.3) and a systematic limiting method to construct rational monopole frames with arbitrary poles and multiplicities:

Theorem 11.5 ([12]). *Given $\alpha_i \in \mathbb{C} \setminus \mathbb{R}$ and positive integers n_i for $1 \le i \le k$, there are soliton monopole frames that are rational and have poles at $\alpha_1, \ldots, \alpha_k$ with multiplicities $n_1, \ldots, n_k$.*

Below is a more general Bäcklund transformation that adds a multiplicity k pole at $\lambda = \alpha$ to a given monopole frame.

Theorem 11.6 ([12]). *Suppose ψ is a monopole frame that is holomorphic and non-degenerate at $\lambda = \alpha$ and ϕ is a soliton (rational) monopole frame with a single pole at $\lambda = \alpha$ with multiplicity k. Then there exist unique $\tilde{\phi}$, $\tilde{\psi}$ such that $\psi_1 = \tilde{\phi}\psi = \tilde{\psi}\phi$ is a monopole frame, $\tilde{\phi}$ is rational with a single pole at $\lambda = \alpha$ with multiplicity k, and $\tilde{\psi}$ is holomorphic and non-degenerate at $\lambda = \alpha$. We use $\phi * \psi$ to denote ψ_1.*

Theorem 11.7 ([13]). *If ψ is a monopole frame with both continuous scattering data and finitely many poles, then there exist unique monopole frames ψ_c and ϕ such that ψ_c has only continuous scattering data, ϕ has only discrete scattering data, and $\psi = \phi * \psi_c$.*

References

[1] Ablowitz, M. J., Kaup, D. J., Newell, A. C., and Segur, H., The inverse scattering transform-Fourier analysis for nonlinear problems. *Studies in Appl. Math.* **53** (1974), 249–315.

[2] Ablowitz, M. J., Clarkson, P. A., *Solitons, non-linear evolution equations and inverse scattering*. London Math. Soc. Lecture Note Ser. 149, Cambridge University Press, Cambridge 1991.

[3] Anand, C. K., Ward's solitons. *Geom. Topol.* **1** (1997), 9–20.

[4] Beals, R., Coifman, R. R., Scattering and inverse scattering for first order systems. *Comm. Pure Appl. Math.* **37** (1984), 39–90.

[5] Beals, R., Coifman, R. R., Inverse scattering and evolution equations. *Comm. Pure Appl. Math.* **38** (1985), 29–42.

[6] Beals, R., Coifman, R. R., Multidimensional inverse scattering and nonlinear partial differential equations. In *Pseudodifferential operators and applications*, Proc. Sympos. Pure Math. 43, Amer. Math. Soc., Providence, RI, 1985, 45–70

[7] Beals, R., Coifman, R. R., Linear spectral problems, non-linear equations and the $\bar{\partial}$-method. *Inverse Problems* **5** (1989), 87–130

[8] Belavin, A. A., Zakharov, V. E., Yang-Mills equations as an inverse scattering problem. *Phys. Lett. B* **73** (1978), 53–57

[9] Brück, M., Du, X., Park, J., and Terng, C. L., The submanifold geometries associated to Grassmannian systems. *Mem. Amer. Math. Soc.* **735** (2002), 1–95.

[10] Burstall, F., Isothermic surfaces: conformal geometry, Clifford algebras and integrable systems. In *Integrable systems, Geometry and Topology*, AMS/IP Stud. Adv. Math. 36, International Press, Cambridge, MA, 2006, 1–82.

[11] Cieśliński, J., Goldstein, P., and Sym, A., Isothermic surfaces in E^3 as soliton surfaces. *Phys. Lett. A* **205** (1995), 37–43.

[12] Dai, B., and Terng, C. L., Bäcklund transformation, Ward solitons, and unitons. arXiv:math.DG/0405363.

[13] Dai, B., Terng, C. L., and Uhlenbeck, K., On the space-time monopole equations. arXiv:math.DG/0602607

[14] Faddeev, L. D., Takhtajan, L. A., *Hamiltonian methods in the theory of solitons*. Springer Ser. Soviet Math., Springer-Verlag, Berlin 1987.

[15] Ferus, D., Pedit, F., Isometric immersions of space forms and soliton theory. *Math. Ann.* **305** (1996), 329–342.

[16] Fokas, A. S. and Ioannidou, T. A., The inverse spectral theory for the Ward equation and for the 2 + 1 chiral model. *Commun. Appl. Anal.* **5** (2) (2001), 235–246.

[17] Gardner, C. S., Greene, J. M., Kruskal, M. D., Miura, R. M., Method for solving the Korteweg-de Vries equation. *Phys. Rev. Lett.* **19** (1967), 1095–1097.

[18] Harnad, J., Saint-Aubin, Y., Shnider, S., The Soliton Correlation Matrix and the Reduction Problem for Integrable Systems. *Comm. Math. Phys.* **93** (1984), 33–56.

[19] Ioannidou, T., Soliton solutions and nontrivial scattering in an integrable chiral model in (2 + 1) dimensions. *J. Math. Phys.* **37** (1996), 3422–3441.

[20] Ioannidou, T., and Zakrzewski, W., Solutions of the modified chiral model in (2 + 1) dimensions. *J. Math. Phys.* **39** (5) (1998), 2693–2701.

[21] Manakov, S. V., and Zakharov, V. E., Three-dimensional model of relativistic-invariant theory, integrable by the inverse scattering transform. *Lett. Math. Phys.* **5** (1981), 247–253.

[22] Pressley, A. and Segal, G., *Loop groups*. Oxford Math. Monogr., Oxford University Press, New York 1986.

[23] Sattinger, D. H., Hamiltonian hierarchies on semi-simple Lie algebras. *Stud. Appl. Math.* **72** (1984), 65–86.

[24] Tenenblat, K., *Transformations of manifolds and applications to differential equations*. Pitman Monogr. Surveys Pure Appl. Math. 93, Longman, Harlow 1998.

[25] Terng, C. L., Soliton equations and differential geometry. *J. Differential Geom.* **45** (1997), 407–445.

[26] Terng, C. L., Geometries and symmetries of soliton equations and integrable elliptic systems. In *Surveys on Geometry and Integrable Systems*, Adv. Stud. in Pure Math., Mathematical Society of Japan, to appear; math.DG/0212372

[27] Terng, C. L., and Uhlenbeck, K., Poisson actions and scattering theory for integrable systems. In *Surveys in differential geometry: integrable systems*, Surv. Diff. Geom. IV, International Press, Boston, MA, 1998, 315–402.

[28] Terng, C. L., Uhlenbeck, K., Schrödinger flows on Grassmannians. In *Integrable systems, Geometry and Topology*, AMS/IP Stud. Adv. Math. 36, International Press, Cambridge, MA, 2006, 235–256.

[29] Terng, C. L., and Uhlenbeck, K., Bäcklund transformations and loop group actions. *Comm. Pure Appl. Math.* **53** (2000), 1–75.

[30] Uhlenbeck, K., Harmonic maps into Lie groups (classical solutions of the chiral model). *J. Differential Geom.* **30** (1989), 1–50.

[31] Uhlenbeck, K., On the connection between harmonic maps and the self-dual Yang-Mills and the sine-Gordon equations. *J. Geom. Phys.* **8** (1992), 283–316.

[32] Villarroel, J., The inverse problem for Ward's system. *Stud. Appl. Math.* **83** (1990), 211–222.

[33] Ward, R. S., Soliton solutions in an integrable chiral model in 2 + 1 dimensions. *J. Math. Phys.* **29** (1988), 386–389.

[34] Ward, R. S., Classical solutions of the chiral model, unitons, and holomorphic vector bundles. *Comm. Math. Phys.* **128** (1990), 319–332.

[35] Ward, R. S., Nontrivial scattering of localized solutions in a $(2+1)$-dimensional integrable systems. *Phys. Letter A* **208** (1995), 203–208.

[36] Zakharov, V. E., Manakov, S. V., The theory of resonant interaction of wave packets in non-linear media. *Soviet Physics JETP* **42** (1974), 842–850.

[37] Zakharov, V. E., and Mikhailov, A. V., Relativistically invariant two dimensional models of fields theory which are integrable by means of the inverse scattering problem method. *Soviet Physics JETP* **47** (6) (1978), 1017–1027.

[38] Zakharov, V. E., Shabat, A. B., Exact theory of two-dimensional self-focusing and one-dimensional of waves in nonlinear media. *Soviet Physics JETP* **34** (1972), 62–69.

[39] Zakharov, V. E., Shabat, A. B., Integration of non-linear equations of mathematical physics by the inverse scattering method, II. *Funct. Anal. Appl.* **13** (1979), 166–174.

Department of Mathematics, University of California at Irvine, Irvine, CA 92697-3875, U.S.A.

E-mail: cterng@math.uci.edu

Finiteness of arithmetic Kleinian reflection groups

Ian Agol*

Abstract. We prove that there are only finitely many arithmetic Kleinian maximal reflection groups.

Mathematics Subject Classification (2000). Primary 30F40; Secondary 57M.

Keywords. Kleinian group, reflection group.

1. Introduction

A Kleinian reflection group Γ is a discrete group generated by reflections in the faces of hyperbolic polyhedron $P \subset \mathbb{H}^3$. We may assume that the dihedral angles of P are of the form π/n, $n \geq 2$, in which case P forms a fundamental domain for the action of Γ on $\mathbb{H}^3$. If P has finite volume, then $\mathbb{H}^3/\Gamma = \mathcal{O}$ is a hyperbolic orbifold of finite volume, which is obtained by "mirroring" the faces of P. Andreev gave a combinatorial characterization of hyperbolic reflection groups in 3-dimensions, in terms of the topological type of P and the dihedral angles assigned to the edges of P [1]. A reflection group Γ is maximal if there is no reflection group Γ' such that $\Gamma < \Gamma'$. We shall defer the definition of arithmetic groups until later, but a theorem of Margulis implies that Γ is arithmetic if and only if $[\mathrm{Comm}(\Gamma) : \Gamma] = \infty$, where $\mathrm{Comm}(\Gamma) = \{g \in \mathrm{Isom}(\mathbb{H}^3) \mid [\Gamma : g^{-1}\Gamma g \cap \Gamma] < \infty\}$ [14]. The main theorem of this paper is that there are only finitely many arithmetic Kleinian groups which are maximal reflection groups.

The argument generalizes an argument of Long–MacLachlan–Reid [12], which implies that there are only finitely many arithmetic minimal hyperbolic 2-orbifolds with bounded genus. Their argument is in fact a generalization of an argument of Zograf [26], who reproved that there are only finitely many congruence groups Γ commensurable with $\mathrm{PSL}(2, \mathbb{Z})$ such that $\mathbb{H}^2/\Gamma$ has genus 0 (this was proven originally by Dennin [4], [5], and was known as Rademacher's conjecture). The key ingredient of their argument is a theorem of Vigneras [20] (based on work of Gelbart–Jacquet [7] and Jacquet–Langlands [10]), which implies that a congruence arithmetic Fuchsian group has $\lambda_1 \geq \frac{3}{16}$ (and which has a generalization to higher dimensions [3]). The other key ingredient is an estimate of Zograf [27], which implies that for a hyperbolic 2-orbifold $\mathcal{O}$, $\lambda_1(\mathcal{O})\mathrm{Vol}(\mathcal{O})$ is bounded linearly by the genus of $\mathcal{O}$. Zograf's argument

*Partially supported by NSF grant DMS-0204142 and the Sloan Foundation.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

is based on a sequence of improvements of a result of Szëgo (who did this for planar domains)[18], by Hersch (for S^2) [9], Yang–Yau (for orientable surfaces) [25], and Li–Yau (for non-orientable surfaces and manifolds of a fixed conformal type) [11]. We observe that the Li–Yau argument (sharpened by El Soufi–Ilias [6]) generalizes to orbifolds, and we then apply this to arithmetic Kleinian reflection groups.

In the concluding section, we consider how one might generalize this result to higher dimensions to prove

Conjecture 1.1. There are only finitely many maximal arithmetic reflection groups in $\mathrm{Isom}(\mathbb{H}^n)$, $n > 1$.

2. Conformal volume of orbifolds

Conformal volume was first defined by Li–Yau, partially motivated by generalizing results on surfaces due to Yang–Yau, Hersch, and Szëgo. We generalize this notion to orbifolds. Let $(\mathcal{O}, g)$ be a compact Riemannian orbifold, possibly with boundary. Let $|\mathcal{O}|$ denote the underlying topological space. Denote the volume form by dv_g, and $\mathrm{Vol}(\mathcal{O}, g)$ its volume. Let $\mathrm{Möb}(\mathbb{S}^n)$ denote the conformal transformations of $\mathbb{S}^n$. It is well known that $\mathrm{Möb}(\mathbb{S}^n) = \mathrm{Isom}(\mathbb{H}^{n+1})$. $|\mathcal{O}|$ has a codimension zero dense open subset which is a Riemannian manifold. We will say that a map $\varphi\colon |\mathcal{O}_1| \to |\mathcal{O}_2|$ is *PC* if it is a continuous map which is piecewise conformal on the open submanifold of $|\mathcal{O}_1|$ which maps to the manifold part of $|\mathcal{O}_2|$. Clearly, if $\varphi\colon |\mathcal{O}| \to \mathbb{S}^n$ is PC, and $\mu \in \mathrm{Möb}(\mathbb{S}^n)$, then $\mu \circ \varphi$ is also PC. Let *can* denote the canonical round metric on $\mathbb{S}^n$.

Definition 2.1. For a piecewise smooth map $\varphi\colon |\mathcal{O}| \to (\mathbb{S}^n, can)$, define

$$V_c(n, \varphi) = \sup_{\mu \in \mathrm{Möb}(\mathbb{S}^n)} \mathrm{Vol}(\mathcal{O}, (\mu \circ \varphi)^*(can)).$$

If there exists a PC map $\varphi\colon |\mathcal{O}| \to \mathbb{S}^n$, then we also define

$$V_c(n, \mathcal{O}) = \inf_{\varphi\colon \mathcal{O} \to \mathbb{S}^n \text{ PC}} V_c(n, \varphi).$$

$V_c(n, \mathcal{O})$ is denoted the (n-dimensional) *conformal volume* of $\mathcal{O}$.

Remark. For our application, it would suffice to define a PC map to be Lipschitz and *a.e.* conformal. It seems likely that our definition of conformal volume coincides with that of Li–Yau for manifolds, but we have not checked this (it would suffice to show that a PC map can be approximated by conformal maps).

If there exists a piecewise isometric map $\varphi\colon (|\mathcal{O}|, g) \to \mathbb{E}^n$ for some n, then clearly $V_c(n, \mathcal{O})$ is well-defined, since $\mathbb{E}^n$ has a conformal embedding into $\mathbb{S}^n$. For an orbifold $(\mathcal{O}, g)$, to prove that $V_c(n, \mathcal{O})$ is well-defined, one need only check that $(|\mathcal{O}|, g)$ embeds piecewise isometrically into some compact Riemannian manifold, in which case the Nash isometric embedding theorem implies that $(|\mathcal{O}|, g)$ embeds

piecewise isometrically into $\mathbb{E}^n$, for some n. We will only apply conformal volume to orbifolds which will obviously have a PC map to $\mathbb{S}^n$, for some n. We record basic facts about conformal volume which were recorded in [11], and which carry over to our notion of conformal volume for orbifolds.

Fact 1. If $\mathcal{O}$ admits a degree d PC map onto another orbifold $\mathcal{P}$, then

$$V_c(n, \mathcal{O}) \leq |d| V_c(n, \mathcal{P}).$$

Fact 2. Since $\mathbb{S}^n \hookrightarrow \mathbb{S}^{n+1}$ embeds isometrically, it is clear that $V_c(n, \mathcal{O}) \geq V_c(n+1, \mathcal{O})$. Define the conformal volume $V_c(\mathcal{O}) = \lim_{n\to\infty} V_c(n, \mathcal{O})$.

Fact 3. If $\mathcal{O}$ is of dimension m, and $\varphi\colon |\mathcal{O}| \to \mathbb{S}^n$ is a PC map, then

$$V_c(n, \varphi) \geq V_c(n, \mathbb{S}^m) = \mathrm{Vol}(\mathbb{S}^m).$$

The same argument as in Li–Yau works here: "blow up" about a smooth manifold point of $\varphi(|\mathcal{O}|)$ so that the image Hausdorff limits to a geodesic sphere of dimension m.

Fact 4. If $\mathcal{O}$ is an embedded suborbifold of the orbifold $\mathcal{P}$, and $\varphi\colon |\mathcal{P}| \to \mathbb{S}^n$ is PC, then $V_c(n, \varphi) \geq V_c(n, \varphi|_{|\mathcal{O}|})$. Thus, $V_c(n, \mathcal{P}) \geq V_c(n, \mathcal{O})$.

3. Finite subgroups of O(3)

Lemma 3.1. *Let $G < \mathrm{O}(3)$ be a finite subgroup. Then there exists a group G', $G \leq G' < \mathrm{O}(3)$, which is generated by reflections such that $[G' : G] \leq 4$.*

Proof. This follows from a case-by-case analysis of spherical 2-orbifolds. We will use Conway's notation for spherical orbifolds (see *e.g.* ch. 13 [19]).

Case $(*)$, $(*p, p)$ or $(*p, q, r)$: These orbifold groups are generated by reflections.

Case (p, p) or (p, q, r): These 2-fold cover $(*p, p)$ or $(*p, q, r)$ respectively.

Case $(n*)$: This 2-fold covers $(*n, 2, 2)$.

Case $(2 * m)$: This 2-fold covers $(*2m, 2, 2)$.

Case $(3 * 2)$: This 2-fold covers $(*4, 3, 2)$.

Case $(n|\circ)$: This 2-fold covers $(2n*) \xrightarrow{2:1} (*2n, 2, 2)$ (this includes the case $n = 1$, *i.e.* $(1|\circ) = \mathbb{RP}^2$).

This exhausts all possible spherical orbifolds, and we see that in each case the orbifold fundamental group is of index ≤ 4 in a reflection group (all but the last case have index ≤ 2). □

Question. Given a dimension n, is there a constant $C(n)$ such that any finite subgroup of $\mathrm{O}(n)$ is of index $\leq C(n)$ in a reflection group? If so, then one should be able to prove Conjecture 7.1. We suspect the answer to this question is no, which is why we have not been able to generalize the proof in this section to higher dimensions.

4. Eigenvalue bounds

We observe that the argument of Thm. 2.2 [6] generalizes to our context of orbifolds (their theorem sharpens Cor. 3, Sect. 2 of [11]). If $(\mathcal{O}, g)$ is a Riemannian orbifold with piecewise smooth boundary, then $\lambda_1(\mathcal{O}, g)$ is the first non-zero eigenvalue of Δ_g on $\mathcal{O}$ with Neumann boundary conditions.

Theorem 4.1 ([6]). *Let $(\mathcal{O}, g)$ be a compact Riemannian orbifold of dimension m, possibly with boundary. If $\varphi\colon |\mathcal{O}| \to \mathbb{S}^n$ is a PC map,*

$$\lambda_1(\mathcal{O}, g)\mathrm{Vol}(\mathcal{O}, g)^{\frac{2}{m}} \leq m V_c(n, \varphi)^{\frac{2}{m}}.$$

Proof. Let $\boldsymbol{X} = (X_1, \ldots, X_{n+1})$ be the coordinate functions on $\mathbb{S}^n \subset \mathbb{R}^{n+1}$. Then $\sum_{i=1}^{n+1} X_i^2 = 1$, restricted to $\mathbb{S}^n$.

Lemma 4.2. *There exists $\mu \in \mathrm{M\ddot{o}b}(\mathbb{S}^n)$ such that $\int_{\mathcal{O}} \boldsymbol{X} \circ \mu \circ \varphi \, dv_g = \mathbf{0}$.*

Proof. Let $\boldsymbol{x} \in \mathbb{S}^n$. For $0 \leq t < 1$, $t\boldsymbol{x} \in \mathbb{H}^{n+1}$, let $\mu_{t\boldsymbol{x}} \in \mathrm{M\ddot{o}b}(\mathbb{S}^n) = \mathrm{Isom}(\mathbb{H}^{n+1})$ be the hyperbolic translation along the ray $\mathbb{R}\boldsymbol{x}$ taking $\mathbf{0}$ to $t\boldsymbol{x}$ (thus, $\mu_{0\mathbf{x}} = \mu_{\mathbf{0}} = \mathrm{Id}$). Let $H(t\boldsymbol{x}) = \frac{1}{\mathrm{Vol}(\mathcal{O},g)} \int_{\mathcal{O}} \boldsymbol{X} \circ \mu_{t\boldsymbol{x}} \circ \varphi \, dv_g$. We may think of H as defining a function $H\colon \mathbb{H}^{n+1} \to \mathbb{H}^{n+1}$, which gives the center of mass of the measure coming from $\varphi_* dv_g$, where we take the point $-t\boldsymbol{x}$ to be the origin of the sphere $\mathbb{S}^n = \partial\mathbb{H}^{n+1}$ by the conformal map $\mu_{t\boldsymbol{x}}$. As $t \to 1$, all of the mass of $\mu_{t\mathbf{x}} \circ \varphi(\mathcal{O})$ becomes concentrated at $\boldsymbol{x}$, and we see that H extends to a continuous function $H\colon B^{n+1} \to B^{n+1}$ (where $B^{n+1} = \mathbb{H}^{n+1} \cup \mathbb{S}^n$) such that $H_{|\mathbb{S}^n} = \mathrm{Id}_{|\mathbb{S}^n}$. Thus, H is onto, so there exists $\boldsymbol{y} \in \mathbb{H}^{n+1}$ such that $H(\boldsymbol{y}) = \mathbf{0}$, and we take $\mu = \mu_{\boldsymbol{y}}$. □

Now replace φ with $\mu \circ \varphi$, noting that this is still a PC map. Then $X_i \circ \varphi$ may be used as test functions in the Rayleigh–Ritz quotient, since they are Lipschitz functions which are L^2 orthogonal to the constant function. Thus,

$$\lambda_1(\mathcal{O}) \int_{\mathcal{O}} |X_i \circ \varphi|^2 \, dv_g \leq \int_{\mathcal{O}} |\nabla X_i \circ \varphi|^2 \, dv_g.$$

Summing, we see that

$$\begin{aligned}
\lambda_1(\mathcal{O}) \int_{\mathcal{O}} \sum_{i=1}^{n+1} |X_i \circ \varphi|^2 \, dv_g &= \lambda_1(\mathcal{O})\mathrm{Vol}(\mathcal{O}, g) \\
&\leq m \int_{\mathcal{O}} \frac{1}{m} \sum_{i=1}^{n+1} |\nabla X_i \circ \varphi|^2 \, dv_g \\
&\leq m \left(\int_{\mathcal{O}} \left(\frac{1}{m} \sum_{i=1}^{n+1} |\nabla X_i \circ \varphi|^2 \right)^{\frac{m}{2}} dv_g \right)^{\frac{2}{m}} \mathrm{Vol}(\mathcal{O}, g)^{1-\frac{2}{m}},
\end{aligned}$$

where the last inequality is Hölder's inequality. Now we use the fact that φ is PC to see that $\varphi^* can = \frac{1}{m}\sum_{i=1}^{n+1} |\nabla X_i \circ \varphi|^2$ *a.e.*, and thus

$$\int_{\mathcal{O}} \Big(\frac{1}{m}\sum_{i=1}^{n+1} |\nabla X_i \circ \varphi|^2\Big)^{\frac{m}{2}} dv_g \leq \mathrm{Vol}(\mathcal{O}, \varphi^* can).$$

Finally, we obtain the desired inequality putting these inequalities together. □

5. Congruence arithmetic hyperbolic 3-orbifolds

We need to know some properties of arithmetic hyperbolic 3-orbifolds. For background and notation, see Maclachlan–Reid [13].

Let $k \subset \mathbb{C}$ be a number field with only one complex place, and let A be a quaternion algebra over k. Let $\rho\colon A \to M(2,\mathbb{C})$ be a k-embedding, $P\colon \mathrm{GL}(2,\mathbb{C}) \to \mathrm{PGL}(2,\mathbb{C})$. Let $\mathcal{E} \subset A$ be either a maximal order or an Eichler order, and let $N(\mathcal{E}) \subset A^*$ be the normalizer of $\mathcal{E}$ in A. Let Γ be an arithmetic Kleinian group, such that $\Gamma = P\rho G$, $G \subset A^*$. Then there exists an order $\mathcal{E} \subset A$ which is either a maximal order or an Eichler order such that $G \subset N(\mathcal{E})$ (see thm. 11.4.3 [13]). In particular, if Γ is a maximal arithmetic Kleinian group, then $\Gamma = P\rho N(\mathcal{E})$, for some order $\mathcal{E}$.

An ideal I in A is a complete R_k-lattice. The left order of I is $\mathcal{O}_l(I) = \{a \in A \mid aI \subset I\}$, and the right order is $\mathcal{O}_r(I) = \{a \in A \mid Ia \subset I\}$. The ideal I is 2-*sided* if $\mathcal{O}_l(I) = \mathcal{O}_r(I)$. The ideal is *integral* if I lies in both $\mathcal{O}_l(I)$ and in $\mathcal{O}_r(I)$ (*i.e.* $I^2 \subseteq I$ is multiplicatively closed). If $\mathcal{O}$ is a maximal order in A, and I is a 2-sided integral ideal in $\mathcal{O}$ such that $\mathcal{O} = \mathcal{O}_l(I) = \mathcal{O}_r(I)$, then the *principal congruence subgroup* of $\mathcal{O}^1$ is

$$\mathcal{O}^1(I) = \mathcal{O}^1 \cap (1 + I).$$

Thus, $\mathcal{O}^1(I)$ is the kernel of the map $\mathcal{O}^1 \to \mathcal{O}/I$, which is therefore of finite index in $\mathcal{O}^1$, since $\mathcal{O}/I$ is finite. A discrete group $G \subset A$ is *congruence* if it contains a principal congruence subgroup, and $\Gamma < \mathrm{PGL}(2,\mathbb{C})$ is congruence if $\Gamma = P\rho G$, for some $G < A$ congruence.

Lemma 5.1 (Long–MacLachlan–Reid [12]). *A maximal arithmetic Kleinian group is congruence.*

Proof. Let $\Gamma \subset \mathrm{PGL}(2,\mathbb{C})$ be a maximal arithmetic Kleinian group. Then $\Gamma = P\rho N(\mathcal{E})$, for some order $\mathcal{E}$ of square-free level. If $\mathcal{E}$ is a maximal order, then $\mathcal{E}^1$ is a congruence subgroup for the trivial ideal $I = \mathcal{E}$. If $\mathcal{E} = \mathcal{O}_1 \cap \mathcal{O}_2$, where $\mathcal{O}_i$ are maximal orders (so that $\mathcal{E}$ is an Eichler order), then choose $\alpha \in R_k - \{0\}$ such that $I = \alpha\mathcal{O}_1 \subset \mathcal{O}_2$. Then $\mathcal{O}_l(I) = \mathcal{O}_r(I) = \mathcal{O}_1$, and $I^2 = \alpha^2\mathcal{O}_1 \subset \alpha\mathcal{O}_1 = I$. Also, clearly $1 + I \subset \mathcal{O}_1 \cap \mathcal{O}_2$. Thus, $\mathcal{O}^1(I) = \mathcal{O}_1^1 \cap (1 + I) \subset \mathcal{O}_1 \cap \mathcal{O}_2$. Thus, $\mathcal{O}^1(I) \subset \mathcal{E}^1$, and we see that $\mathcal{E}^1$ is a principal congruence subgroup. □

A fundamental theorem of Vigneras, making use of work of Jacquet–Langlands [10] and Gelbart–Jacquet [7], generalizes a result of Selberg for $\mathrm{PSL}(2,\mathbb{Z})$. For a hyperbolic orbifold $\mathcal{O}$, let $\lambda_1(\mathcal{O})$ be the minimal non-zero eigenvalue of the Laplacian Δ on $\mathcal{O}$.

Theorem 5.2 ([20], [3]). *Let $\mathcal{O} = \mathbb{H}^3/\Gamma$, where Γ is an arithmetic congruence subgroup. Then $\lambda_1(\mathcal{O}) \geq \frac{3}{4}$.*

It is conjectured that under the hypotheses of the above theorem, $\lambda_1(\mathcal{O}) \geq 1$, which is known as (a special case of) the *generalized Ramanujan conjecture* [3].

6. Finiteness of arithmetic Kleinian maximal reflection groups

We put together the results from the previous sections to prove our main theorem.

Theorem 6.1. *There are only finitely many arithmetic maximal reflection groups in* $\mathrm{Isom}(\mathbb{H}^3)$.

Proof. Suppose that Γ is an arithmetic maximal reflection group. That is, there is no group $\Gamma' < \mathrm{Isom}(\mathbb{H}^3)$, with $\Gamma < \Gamma'$, such that Γ' is generated by reflections. Then there exists $\Gamma \leq \Gamma_0 < \mathrm{Isom}(\mathbb{H}^3)$, such that Γ_0 is a maximal Kleinian group. Γ is generated by reflections in a finite volume polyhedron P.

Lemma 6.2 (Vinberg [21]). *Γ is a normal subgroup of Γ_0. Moreover, there is a finite subgroup $\Theta < \Gamma_0$ such that $\Theta \to \Gamma_0/\Gamma$ is an isomorphism, and Θ preserves the polyhedron P.*

Proof. This follows from the fact that the set of reflections in Γ_0 is conjugacy invariant, and therefore the group generated by reflections is normal in Γ_0. Since Γ is a maximal reflection group, this subgroup must be Γ. The polyhedron P is a fundamental domain of Γ, and if there is an element $\gamma \in \Gamma_0$ such that $\mathrm{int(P)} \cap \gamma(\mathrm{int(P)}) \neq \emptyset$, then $\gamma(P) = P$. Otherwise, there would be a geodesic plane V containing a face of P, such that $V \cap \mathrm{int(P)} \neq \emptyset$. The reflection $r_V \in \Gamma$ in the plane V would be conjugate to a reflection $r_{\gamma(V)} = \gamma r_V \gamma^{-1}$, which is not in Γ since $r_{\gamma(V)}(\mathrm{int(P)}) \cap \mathrm{int(P)} \neq \emptyset$, which implies that Γ is not a maximal reflection group, a contradiction. Let Θ be the subgroup of Γ_0 such that $\Theta(P) = P$. Clearly Θ is finite, since P is finite volume and has finitely many faces. If $\gamma_0 \in \Gamma_0$, let $\gamma \in \Gamma$ be such that $\gamma_0(\mathrm{int(P)}) \cap \gamma(\mathrm{int(P)}) \neq \emptyset$. Then $\gamma^{-1}\gamma_0(P) = P$, so $\gamma_0 \in \gamma\Theta$. Thus, $\Theta \to \Gamma_0/\Gamma$ is an isomorphism. □

Let $\mathcal{O} = \mathbb{H}^3/\Gamma_0$, and Θ is the finite group coming from the previous lemma.

Lemma 6.3. $\lambda_1(\mathcal{O}) = \lambda_1(P/\Theta)$.

Proof. Let f be an eigenfunction on P/Θ with eigenvalue $\lambda_1(P/\Theta)$. Since f has Neumann boundary conditions, its level sets in P/Θ are orthogonal to $\partial P/\Theta$. Let $\tilde{f}$

be the preimage of f under the map $P \to P/\Theta$, so that $\tilde{f}$ is invariant under the action of Θ. Extend $\tilde{f}$ to a function $\tilde{F}$ on $\mathbb{H}^3$, by the action of Γ (and therefore invariant under the action of Γ_0). By the reflection principle, $\tilde{F}$ is a smooth function, so it descends to an eigenfunction F of Δ on $\mathcal{O}$ with eigenvalue $\lambda_1(P/\Theta)$. Conversely, if F is an eigenfunction of Δ on $\mathcal{O}$ with eigenvalue $\lambda_1(\mathcal{O})$, then $F_{|P/\Theta}$ gives an eigenfunction on P/Θ with Neumann boundary conditions, since the level sets of $\tilde{F}$ must be invariant under reflections, and therefore perpendicular to the faces of P. □

Consider $\mathbb{H}^3 \subset \mathbb{S}^3$ embedded conformally as the upper half space of $\mathbb{S}^3$, so that $\mathrm{Isom}(\mathbb{H}^3)$ acts conformally on $\mathbb{S}^3$. Normalize so that Θ acts isometrically on $\mathbb{S}^3$, which we may do by a result of Wilker [24]. Clearly, $V_c(n, \mathcal{O}) = V_c(n, P/\Theta)$, since $|\mathcal{O}| = |P/\Theta|$. Then the orbifold $P/\Theta \subset \mathbb{S}^3/\Theta$ is a conformal embedding, so by Fact 4, $V_c(3, P/\Theta) \leq V_c(3, \mathbb{S}^3/\Theta)$. The group Θ embeds in a finite reflection group $\Theta' \subset \mathrm{O}(3)$ such that $[\Theta' : \Theta] \leq 4$. Then by Fact 1, $V_c(3, \mathbb{S}^3/\Theta) \leq 4V_c(3, \mathbb{S}^3/\Theta')$. Now, there is a polyhedron $Q \subset \mathbb{S}^3$ with geodesic faces which is the fundamental domain for Θ'. Clearly $V_c(3, Q) = V_c(3, \mathbb{S}^3/\Theta')$. By Facts 2 and 4, $V_c(3, Q) = \mathrm{Vol}(\mathbb{S}^3, \mathit{can}) = 2\pi^2$. Thus, we have $V_c(\mathcal{O}) \leq 8\pi^2$.

Now we apply the eigenvalue estimates

$$\frac{3}{4}\mathrm{Vol}(\mathcal{O})^{\frac{2}{3}} \leq \lambda_1(\mathcal{O})\mathrm{Vol}(\mathcal{O})^{\frac{2}{3}} \leq 3V_c(\mathbb{H}^3/\Gamma_0)^{\frac{2}{3}} \leq 3(8\pi^2)^{\frac{2}{3}}.$$

Thus we obtain $\mathrm{Vol}(\mathcal{O}) \leq 64\pi^2$. Since volumes of arithmetic hyperbolic orbifolds are discrete, and Γ_0 is uniquely determined by Γ, we conclude that there are only finitely many arithmetic maximal reflection groups. □

7. Conclusion

From the main theorem, we conclude that given an arithmetic reflection group $\Gamma < \mathrm{Isom}(\mathbb{H}^3)$, it must lie in one of finitely many maximal reflection groups (up to conjugacy). If Γ is a reflection group in a polyhedron P for which all the dihedral angles are $\pi/2$ or all are $\pi/3$, then there are infinitely many co-finite volume reflection subgroups of Γ. Thus we see that there are commensurability classes of arithmetic groups for which there are infinitely many reflection groups in the commensurability class, and thus in our finiteness result, the maximality assumption is crucial. It is an interesting project to try to identify all arithmetic maximal reflection groups, and to classify their reflection subgroups of finite covolume.

For the non-compact examples, one may apply volume formulae to estimate the maximal discriminant of a quadratic imaginary number field k for which $\mathrm{PGL}(2, k)$ contains a reflection group. Humbert first computed the covolumes of Bianchi groups, and a generalization due to Borel implies that for a non-compact arithmetic Kleinian

group, the minimal covolume μ satisfies

$$\mu \geq \frac{|\Delta_k|^{\frac{3}{2}}\zeta_k(2)}{16\pi^2 h_k},$$

where h_k is the class number of the number field k. The Brauer–Siegel theorem gives an estimate of h_k for a number field and implies for a quadratic imaginary number field k that

$$|\Delta_k|\zeta_k(2) \geq \frac{h_k(2\pi)^2}{2w},$$

where w is the order of the group of roots of unity in k. If $k \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$, then $w = 2$ (see the proof of theorem 11.7.2 [13]). Thus, we have

$$64\pi^2 \geq \mu \geq \frac{1}{16}|\Delta_k|^{\frac{1}{2}}.$$

Thus $|\Delta_k| < 2^{20}\pi^4 = 1.02 \times 10^8$. Hatcher has computed orbifold structures of some of the Bianchi groups of small discriminant, and from his pictures one may deduce that $\mathrm{PGL}(2, R_k)$ is commensurable with a reflection group for

$$\begin{aligned}\Delta_k = &-3,\ -4,\ -7,\ -8,\ -11,\ -15,\ -19,\ -20,\\ &-24,\ -39,\ -40,\ -52,\ -55,\ -56,\ -68,\ -84,\end{aligned}$$

where k is a quadratic imaginary number field [8]. In principle, it ought to be possible to compute all arithmetic reflection groups, but clearly even in the non-compact case, the volume estimates we obtain do not make this computation feasible. To classify non-compact Kleinian arithmetic groups, it may require the infusion of some more number theory. Arithmetic restrictions have been found on reflection groups containing the Bianchi groups by Vinberg [23] and Blume-Nienhaus [2]. If these results could be extended to all maximal non-compact arithmetic groups, then one may be able to give a complete classification. The classification of compact arithmetic maximal reflection groups, although in principle decidable, is probably not feasible at this stage.

It is clear that for a finite volume polyhedron $P \subset \mathbb{H}^n$, $V_c(n, P) = \mathrm{Vol}(\mathbb{S}^n)$, so

$$\lambda_1(P)\mathrm{Vol}(P)^{\frac{2}{n}} < n\mathrm{Vol}(\mathbb{S}^n)^{\frac{2}{n}}.$$

Thus in n dimensions, there are finitely many reflection groups $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ which have a lower bound on $\lambda_1(\mathbb{H}^n/\Gamma)$. It would be interesting if one could generalize the arguments of the main theorem to higher dimensions. It is known by work of Prokhorov that there cannot be any reflection groups in dimension > 996 [17] (Vinberg showed that compact reflection groups cannot exist in dimension > 30 [22]). Vinberg also gave a characterization of arithmetic reflection groups, in terms of a totally real field of definition K and an integrality condition [21]. Nikulin has shown that there

exists a constant N_0 such that the set of arithmetic groups in $\mathrm{Isom}(\mathbb{H}^n)$ generated by reflections with $n > 16$ and $[K : \mathbb{Q}] > N_0$ is empty [16]. It was proved in a previous paper by Nikulin that the set of maximal arithmetic groups generated by reflections in $\mathrm{Isom}(\mathbb{H}^n)$ with a fixed degree $[K : \mathbb{Q}]$ is finite [15]. Thus, to generalize the main argument of this paper, one would have to bound the conformal volume of a minimal arithmetic n-orbifold containing a reflection suborbifold up to dimensions $n \leq 16$. By the characterization of maximal arithmetic groups in $\mathrm{Isom}(\mathbb{H}^n)$, it should follow that they are congruence. By a result of Burger–Sarnak, if $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ is a congruence arithmetic reflection group with $n > 2$, then $\lambda_1(\mathbb{H}^n/\Gamma) > \frac{2n-3}{4}$ (Cor. 1.3 [3], and the fact that reflection groups are defined by quadratic forms). Every maximal arithmetic group covered by a reflection group will embed conformally in an elliptic n-orbifold. So to prove the following conjecture for $n \leq 16$:

Conjecture 7.1. There is a function $K(n)$, such that if $\mathcal{O}$ is an elliptic n-orbifold, then $V_c(\mathcal{O}) \leq K(n)$.

References

[1] Andreev, E. M., Convex polyhedra in Lobačevskiĭ spaces. *Mat. Sb.* (N.S.) **81 (123)** (1970), 445–478.

[2] Blume-Nienhaus, Jürgen, *Lefschetzzahlen für Galois-Operationen auf der Kohomologie arithmetischer Gruppen*. Bonner Mathematische Schriften 230, Universität Bonn, Mathematisches Institut, Bonn 1992.

[3] Burger, M., and Sarnak, P., Ramanujan duals. II. *Invent. Math.* **106** (1) (1991), 1–11.

[4] Dennin, Joseph B., Jr., Fields of modular functions of genus 0. *Illinois J. Math.* **15** (1971), 442–455.

[5] —, Subfields of $K(2^n)$ of genus 0. *Illinois J. Math.* **16** (1972), 502–518.

[6] Soufi, A. El, and Ilias, S., Immersions minimales, première valeur propre du laplacien et volume conforme. *Math. Ann.* **275** (2) (1986), 257–267.

[7] Gelbart, Stephen, and Jacquet, Hervé, A relation between automorphic representations of GL(2) and GL(3). *Ann. Sci. École Norm. Sup.* (4) **11** (1978), 471–542.

[8] Hatcher, A., Bianchi orbifolds of small discriminant. Preprint, 1983; http://www.math.cornell.edu/~hatcher/bianchi.html.

[9] Hersch, Joseph, Quatre propriétés isopérimétriques de membranes sphériques homogènes. *C. R. Acad. Sci. Paris Sér. A-B* **270** (1970), A1645–A1648.

[10] Jacquet, H., and Langlands, R. P., *Automorphic forms on* GL(2). Lecture Notes in Math. 114, Springer-Verlag, Berlin 1970.

[11] Li, Peter, and Yau, Shing Tung, A new conformal invariant and its applications to the Willmore conjecture and the first eigenvalue of compact surfaces. *Invent. Math.* **69** (2) (1982), 269–291.

[12] Long, D. D., Reid, A. W., and Maclachlan, C., Arithmetic fuchsian groups of genus zero. Preprint, 2005.

[13] Maclachlan, Colin, and Reid, Alan W., *The arithmetic of hyperbolic 3-manifolds*. Grad. Texts in Math. 219, Springer-Verlag, New York 2003.

[14] Margulis, G. A., Discrete groups of motions of manifolds of nonpositive curvature. In *Proceedings of the International Congress of Mathematicians* (Vancouver, B.C., 1974), Vol. 2, Canad. Math. Congress, Montreal, Que., 1975, 21–34.

[15] Nikulin, V. V., On the arithmetic groups generated by reflections in Lobačevskiĭ spaces. *Izv. Akad. Nauk SSSR Ser. Mat.* **44** (3) (1980), 637–669, 719–720.

[16] —, On the classification of arithmetic groups generated by reflections in Lobačevskiĭ spaces. *Izv. Akad. Nauk SSSR Ser. Mat.* **45** (1) (1981), 113–142, 240.

[17] Prokhorov, M. N., Absence of discrete groups of reflections with a noncompact fundamental polyhedron of finite volume in a Lobačevskiĭ space of high dimension. *Izv. Akad. Nauk SSSR Ser. Mat.* **50** (2) (1986), 413–424.

[18] Szegö, G., Inequalities for certain eigenvalues of a membrane of given area. *J. Rational Mech. Anal.* **3** (1954), 343–356.

[19] Thurston, William P., *The geometry and topology of 3-manifolds*. Lecture notes from Princeton University, 1978–80.

[20] Vignéras, Marie-France, Quelques remarques sur la conjecture $\lambda_1 \geq \frac{1}{4}$. In *Seminar on number theory* (Paris, 1981/1982), Progr. Math. 38, Birkhäuser, Boston, MA, 1983, 321–343.

[21] Vinberg, È. B., Discrete groups generated by reflections in Lobačevskiĭ spaces. *Mat. Sb.* (N.S.) **72 (114)** (1967), 471–488; correction, ibid. **73 (115)** (1967), 303.

[22] —, Absence of crystallographic groups of reflections in Lobačevskiĭ spaces of large dimension. *Trudy Moskov. Mat. Obshch.* **47** (1984), 68–102, 246.

[23] —, Reflective subgroups in Bianchi groups. *Selecta Math. Soviet.* **9** (4) (1990), 309–314.

[24] Wilker, J. B., Isometry groups, fixed points and conformal transformations. *C. R. Math. Rep. Acad. Sci. Canada* **4** (5) (1982), 293–297.

[25] Yang, Paul C., and Yau, Shing Tung, Eigenvalues of the Laplacian of compact Riemann surfaces and minimal submanifolds. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **7** (1980), 55–63.

[26] Zograf, P., A spectral proof of Rademacher's conjecture for congruence subgroups of the modular group. *J. Reine Angew. Math.* **414** (1991), 113–116.

[27] Zograf, P. G., Small eigenvalues of automorphic Laplacians in spaces of cusp forms. Automorphic functions and number theory, II. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **134** (1984), 157–168.

Department of Mathematics, Statistics, & Computer Science, University of Illinois at Chicago, 322 Science & Engineering Offices (M/C 249), 851 S. Morgan St., Chicago, IL 60607-7045, U.S.A.

E-mail: agol@math.uic.edu

Non-positive curvature and complexity for finitely presented groups

Martin R. Bridson*

Abstract. A universe of finitely presented groups is sketched and explained, leading to a discussion of the fundamental role that manifestations of non-positive curvature play in group theory. The geometry of the word problem and associated filling invariants are discussed. The subgroup structure of direct products of hyperbolic groups is analysed and a process for encoding diverse phenomena into finitely presented subdirect products is explained. Such an encoding is used to solve problems of Grothendieck concerning profinite completions and representations of groups. In each context, explicit groups are crafted to solve problems of a geometric nature.

Mathematics Subject Classification (2000). Primary 20F65; Secondary 20F67.

Keywords. Geometric group theory, finitely presented groups, non-positive curvature, Dehn functions, filling invariants, decision problems.

Introduction

When viewed through the eyes of a topologist, a finite group-presentation $\Gamma = \langle \mathcal{A} \mid \mathcal{R} \rangle$ is a concise description of a compact, connected, 2-dimensional CW-complex K with one vertex: the generators $a \in \mathcal{A}$ index the (oriented) 1-cells and the defining relations $r \in \mathcal{R}$ describe the loops along which the boundaries of the 2-cells are attached. Γ emerges as the group of deck transformations of the universal cover $\tilde{K}$ and the Cayley graph $\mathcal{C}_{\mathcal{A}}(\Gamma)$ is the 1-skeleton of $\tilde{K}$.

Thus we meet the two main strands of geometric group theory, intertwined as they often are. In the first and most classical strand, one studies actions of groups on metric and topological spaces in order to elucidate the structure of the space and the group. The quality of the insights that one obtains varies with the quality of the action: one may prefer discrete cocompact actions by isometries on spaces with fine geometric structure, but according to context one must vary the conditions on the action, sometimes weakening admission criteria to obtain a more diverse class of groups, sometimes demanding more structure to narrow the focus and study groups and spaces of exceptional character.

This first strand mingles with the second, wherein one regards finitely generated groups as geometric objects in their own right [62], equipped with *word metrics*: given a finite generating set S for a group Γ one defines $d_S(\gamma_1, \gamma_2)$ to be the length of

*This work was supported in part by an EPSRC Advanced Fellowship.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

the shortest word in the free group on S that is equal to $\gamma_1^{-1}\gamma_2$ in Γ. In other words, d_S is the restriction to the vertex set of the standard length metric on the Cayley graph of Γ. The word metric and Cayley graph depend on the choice of generating set, but their quasi-isometry type does not. Thus one is particularly interested in properties of groups and spaces that are invariant under quasi-isometry. When dealing with such invariants, one is free to replace Γ by any space that is quasi-isometric to it, such as the universal cover of a closed Riemannian manifold with fundamental group Γ, where the tools of analysis can be brought to bear.

The techniques of geometric group theory merge into the more combinatorial techniques that held sway in the study of finitely presented groups for most of the twentieth century. At the heart of combinatorial group theory lie the fundamental decision problems first articulated by Max Dehn [50] – the word, conjugacy and isomorphism problems. These continue to play an important role in geometric group theory and provide a unifying theme for the ideas presented here, serving as fundamental measures of the complexity of groups.

In this article and the accompanying lecture I shall discuss two topics that account for much of my work: manifestations of non-positive curvature in group theory, and the geometry of the word problem and associated filling invariants. I shall also explore the subgroup structure of direct products of hyperbolic groups. It transpires that a huge range of phenomena can be encoded into the finitely presented subgroups of such direct products. Fritz Grunewald and I used such an encoding to solve problems of Grothendieck concerning profinite completions and representations of groups; this is explained in Section 6. Throughout the discussion, the reader will find that a prominent role is played by explicit groups crafted to solve problems of a geometric nature.

Acknowledgements. It is a pleasure to thank my coauthors, past and present, for their ideas and companionship. It is also a pleasure to acknowledge the great intellectual debt that I owe to Mikhael Gromov.

1. The universe of finitely presented groups

The picture of the universe of groups that I am about to sketch[1] is shaded by personal taste and the needs of today but nevertheless I claim that it has intrinsic merit. The point of attempting such a sketch is that it forces one to consider how different classes of groups that arise in disparate settings are related; it challenges one to locate given groups in relation to others and to explain how different classes intersect; and it helps to tease out why certain classes are worthy of particular attention. One asks what theorems hold where, and how various measures of complexity (decision problems, subgroup structure, finiteness properties, ...) vary and decay as one moves around the universe. One finds oneself reflecting on how problems from elsewhere in mathematics can be encoded in groups of one type but not another, and one starts to wonder

[1] I thank Tim Riley for his skillful rendering of my hand-drawing.

how such problems might be transported to a more hospitable region of the universe where one has stricter definitions and better theorems to mount an attack.

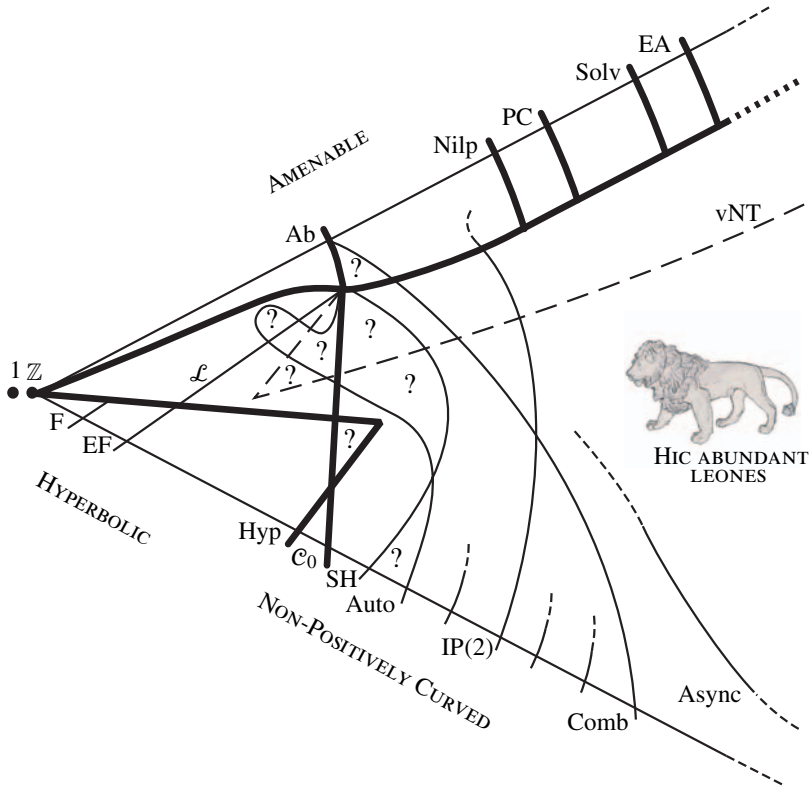


Figure 1. The universe of finitely presented groups.

When approaching group theory from the viewpoint of large-scale geometry, it is natural to blur the distinction between commensurable groups. Thus our universe begins with a single (large and interesting) point labelled 1 representing the finite groups. The simplest infinite group is surely $\mathbb{Z}$, so we have a second point representing the virtually cyclic groups. Here the universe divides. If one wants to retain the safety of commutativity and amenability, one can proceed from $\mathbb{Z}$ to the virtually abelian groups. As one slowly relinquishes commutativity and control over growth and constructibility, one passes through the progressively larger classes of (virtually-) nilpotent, polycyclic, solvable and elementary amenable groups, which are marked in the region bounded by a thick line enclosing the amenable groups.

Thinking more freely, instead of taking direct products one might proceed from $\mathbb{Z}$ by taking free products, moving into the class F of virtually free[2] groups, with their tree-like Cayley graphs. As one proceeds away from F, the infinite curvature of tree-

[2]F contains only one commensurability class besides $\mathbb{Z}$, but is drawn larger for effect.

ness gives way to the strictly negative curvature of hyperbolic groups Hyp, then the increasingly weak forms of non-positive curvature that define the classes discussed in the sections that follow: $\mathcal{C}_0$, the groups that act properly and cocompactly by isometries on CAT(0) spaces; SH, the semihyperbolic groups of [2]; the automatic groups Auto of [55]; and the combable and asynchronously combable groups Comb and Async. The line marked IP(2) encloses the groups that satisfy a quadratic isoperimetric inequality (4.3). The thickness of the line delimiting $\mathcal{C}_0$ expresses the view that these groups deserve particular attention.

The *von Neumann–Tits line* (vNT) separates the groups that contain non-abelian free groups from those that do not. For each group below the line, one asks whether its finitely generated subgroups satisfy a *Tits alternative*: if non-amenable, they should contain a non-abelian free group. One also asks what type of amenable subgroups can arise, cf. (2.3).

In contrast to the other amenable groups, virtually abelian groups are indisputably "non-positively curved". Several classes of non-positively curved groups serve as natural envelopes uniting free and free-abelian groups. These include the right-angled Artin groups, cocompact groups of isometries of CAT(0) cube complexes, and limit groups (marked $\mathcal{L}$). The case for placing the class EF $\subset \mathcal{L}$ of *elementarily free groups* immediately next to F is discussed in (1.2). The fundamental group Σ_g of any closed surface of genus $g \geq 2$ is in EF, lending substance to the tradition in combinatorial group theory that, among non-free groups, it is the Σ_g that resemble free groups most closely (cf. 1.8).

1.1. Accuracy and completeness. Adapting to today's foci, and so as not to crowd the diagram, I have omitted many natural classes of groups, even non-positively curved groups. In particular, I have not subdivided $\mathcal{C}_0$ according to subclasses of CAT(0) spaces such as CAT(-1) spaces, spaces with isolated flats, or the classes mentioned above in connection with $\mathbb{Z}^n$; these all enjoy important additional properties and have a rich hoard of examples.

The hints at subdivision in Comb are intended to suggest the various subclasses defined by varying the degree of control one demands over the geometry and linguistic complexity of the combing, cf. (3.2). A question mark in a region of intersection indicates that it is unknown if that region is empty. The question mark at the amenable end of Comb asks, more generally, which amenable groups are combable. Likewise, the vague manner in which the boundary of IP(2) ends reflects a lack of knowledge about which amenable groups have quadratic Dehn functions (4.3). The extent of Async is also unknown, though we know it contains many solvable groups and the fundamental groups of compact 3-manifolds (3.5).

1.2. Limits and ultralimits. From a geometric perspective (as well as others) the virtually nilpotent groups have an indisputable claim to the ground next to abelian groups: they are exactly the groups of polynomial growth, as Gromov proved in the landmark paper [63]. A key idea in that paper is to construct a space on which a

group Γ of polynomial growth acts by taking the limit in the Gromov–Hausdorff topology of a subsequence of the pointed metric spaces $X_n = (\Gamma, \frac{1}{n}d)$, where d is a fixed word metric on Γ (the identity serves as a basepoint).

To circumvent the failure of the sequence $X_n = (\Gamma, \frac{1}{n}d)$ to converge in the non-nilpotent case, one fixes a non-principal ultrafilter and takes an *ultralimit* to obtain an *asymptotic cone* $\mathrm{Cone}_\omega\Gamma$. Such cones have been intensively studied in recent years, particularly in connection with quasi-isometric rigidity, e.g. [73]. Their geometry and algebraic topology encode a good deal of information about Γ (see [64], [83], [54]). If Γ is a non-abelian free group, then $\mathrm{Cone}_\omega\Gamma$ is an everywhere-branching $\mathbb{R}$-tree (regardless of the choice of ω). The class of finitely generated groups that share this property consists precisely of the non-elementary *hyperbolic groups*. This is just one of the ways in which hyperbolic groups appear as the most commanding generalisation of a free group.

1.3. Next to the free groups: limit groups. Continuing with the idea of taking limits, one might ask which finitely generated groups arise as Gromov–Hausdorff limits of free groups [47]. More precisely[3], given Γ, one asks if there exists a finite generating set S for Γ and a sequence of finite generating sets S_n for a fixed free group F with bijections $S_n \to S$ making the ball of radius $\rho \in \mathbb{N}$ about the identity in the Cayley graph $\mathcal{C}_{S_n}(F)$ isomorphic (as a marked graph) to that in $\mathcal{C}_S(\Gamma)$ whenever $n \geq N(\rho)$.

The groups that arise in this way are the *limit groups* $\mathcal{L}$, which are non-positively curved in every reasonable sense [1]. The fundamental importance of this class has been greatly illuminated in recent years by the work of Zlil Sela and others (see Subsection 5.3). A fascinating aspect of their study is that the same class of groups emerges from a range of different definitions that make precise the idea of an *approximately free group*. In terms of first order logic, $\mathcal{L}$ consists of those finitely generated groups that have the same existential theory as a free group [82], while $\mathrm{EF} \subset \mathcal{L}$ consists of those that have the same elementary theory.

1.4. Beware of the lions: encoding. Constructions such as the Higman Embedding Theorem [68] show that one can encode the workings of an arbitrary Turing machine into a finite group-presentation. The groups that one obtains from such constructions will typically not belong to any of our marked classes but rather will lie in the region where the fierce lion is shown defying the groups that submit to the control of our definitions. One should therefore regard the lion as a warning that it is reckless to base a conjecture about arbitrary finitely presented groups on evidence gained along the coasts of the universe.

1.5. Embracing the lions: subdirect products. It is not satisfactory to content oneself with the study of groups in our marked classes alone. In particular one wants to attack problems from elsewhere in mathematics by exploiting the ability

[3]The corresponding topology on marked groups is often called the Grigorchuk topology [60].

to encode arbitrary recursive phenomena into finite group-presentations. In order to do so without leaving the safety of the regions where one has definitions and theorems, one has to find a way of encoding arbitrary presentations into the structure of groups in the labelled classes (the closer to the origin 1 the better). Such encodings exist in remarkable generality, provided one is prepared to accept passing to finitely presented subgroups in the given class. This is the theme of Section 5 below, where I explain and exploit the fact that finitely presented subgroups of direct products of hyperbolic groups (but not limit groups) can be made extremely complicated. Section 6 shows how such an encoding can be used to solve problems of interest elsewhere in mathematics.

1.6. Special groups. One might begin an open-minded search for groups of special interest by asking what sort of actions are admitted by an arbitrary finitely presented group. As one tries to improve the quality of the space or action, obstructions emerge and special groups are singled out. For example, when one knows that every finitely presented group is the fundamental group of a closed symplectic 4-manifold and a closed complex manifold, it is natural to ask which groups are fundamental groups of compact Kähler manifolds, or of 3-manifolds, and to pay special attention to such groups, classifying them if possible. It is also stimulating to try to locate them in our map of the universe.

1.7. Classifying spaces. Other special classes emerge as one tries to improve on the construction of $K = K(\mathcal{A}, \mathcal{R})$ in the opening paragraph, making it more highly-connected and looking for a classifying space. *Higher finiteness properties* emerge and conditions that ensure the existence of a compact $K(\Gamma, 1)$ come into focus, such as the existence of a 1-relator or small-cancellation presentation. Complete non-positively curved spaces serve as classifying spaces but these are not always of the minimum possible dimension [25], [18]. This discrepancy fits into a large body of work in which different notions of dimension measure the cost of choosing between algebraic, topological and geometric models for a group.

Of the classes in figure 1, hyperbolic, polycyclic and $\mathcal{C}_0$ groups all act properly and cocompactly on contractible, finite-dimensional complexes but certain groups in Solv, IP(2) and Async do not. Members of the classes other than Solv, IP(2) and EA have classifying spaces with only finitely many cells in each dimension.

1.8. The first groups and their automorphisms. The view that $\mathbb{Z}^n$, F_n and Σ_g are the most basic of infinite groups begins a rich vein of ideas concerning the automorphisms of these groups. At the level of individual automorphisms, classical facts about integer matrices are paralleled by the *Nielsen–Thurston theory* of surface automorphisms and, for free groups, the *train-track technology* of Bestvina, Feighn and Handel [10], [11].

The analogies between the outer automorphism groups $\mathrm{GL}(n, \mathbb{Z})$, $\mathrm{Out}(F_n)$ and $\mathrm{Mod}_g \cong \mathrm{Out}(\Sigma_g)$ (the mapping class group) go far beyond the observation that

$\mathrm{GL}(2,\mathbb{Z}) \cong \mathrm{Out}(F_2) \cong \mathrm{Mod}_1$. Indeed much of the work on mapping class groups and automorphisms of free groups is premised on such analogies [9], [91], [44]. Karen Vogtmann has been particularly influential in promoting this idea.

2. CAT(0) spaces and their isometries

The theory of CAT(0) spaces has had a huge impact in recent years, not only in geometric group theory but also in the study of low-dimensional manifolds and rigidity phenomena in geometry. (This influence, which owes much to the work of Gromov, is easy to discern in the proceedings of recent ICMs including this one.) My purpose here is not to survey this field but rather to highlight some features of the basic theory with an eye on their quasi-fication in the next section. My book with André Haefliger [35] provides a thorough introduction to the subject.

2.1. CAT(0) spaces. Following A.D. Alexandrov, one defines non-positive curvature in the context of length spaces X by means of the *CAT*(0) *inequality*, which requires that small triangles in X be no fatter than their comparison triangles in the Euclidean plane. Thus one compares triangles $\Delta = \Delta(x_1, x_2, x_3)$ consisting of three points $x_1, x_2, x_3 \in X$ and three geodesic segments $[x_i, x_j]$ to triangles $\overline{\Delta}(\bar{x}_1, \bar{x}_2, \bar{x}_3) \subset \mathbb{E}^2$ with $d(\bar{x}_i, \bar{x}_j) = d(x_i, x_j)$. A point $\bar{p}$ on the line segment $[\bar{x}_i, \bar{x}_j]$ is called a *comparison point* for $p \in [x_i, x_j]$ if $d(x_i, p) = d(\bar{x}_i, \bar{p})$.

A geodesic space is said to be a CAT(0) *space* if for all triangles Δ in that space, $d(p, q) \leq d(\bar{p}, \bar{q})$ for all comparison points $\bar{p}, \bar{q} \in \overline{\Delta}$. And a metric space X is defined to be *non-positively curved* if every point of X has a neighbourhood that, when equipped with the induced metric, is a CAT(0) space. Similarly, one defines the notion of a CAT(-1) space by taking comparison triangles in the hyperbolic plane, and one defines a space to be negatively curved (in the sense of A. D. Alexandrov) if it is locally CAT(-1).

Because the CAT(0) condition encapsulates the essence of non-positive curvature so well, non-positively curved metric spaces satisfy many of the elegant features inherent in the theory of Riemannian manifolds of non-positive sectional curvature. At the heart of the theory lie local-to-global phenomena that spring from the fact that the metric on a CAT(0) space X is convex: if $c_1, c_2 \colon [0, 1] \to X$ are geodesics, then for all $t \in [0, 1]$

$$d(c_1(t),\, c_2(t)) \leq (1-t)\, d(c_1(0),\, c_2(0)) + t d(c_1(1),\, c_2(1)). \tag{2.1}$$

This inequality implies that there is a unique geodesic segment joining each pair of points in X and that X is contractible. The most important example of a local-to-global phenomenon is the Cartan–Hadamard Theorem: *If a complete, simply-connected metric space is non-positively curved, then it is* CAT(0) *space.* (See Chapter II.4 of [35] for a more general result and references.)

It follows from this theorem that compact non-positively curved spaces have contractible universal covers and hence provide classifying spaces. The usefulness of this observation is greatly enhanced by two facts. First, Gromov's *link condition* ([35], p. 206) enables one to reduce the question of whether a polyhedral complex supports a metric of non-positive curvature to a question about the geometry of links in that complex; this allows arguments that proceed by induction on the dimension of the complex, and if the cells are sufficiently regular (e.g. cubes) it can lead to purely combinatorial criteria for the existence of metrics of non-positive curvature. Secondly, gluing theorems ([35], II.11) allow one to preserve non-positive curvature while combining spaces according to group-theoretic constructions that one wishes to perform at the level of π_1, such as amalgamated free products $\Gamma_1 *_{\mathbb{Z}} \Gamma_2$.

2.2. Splitting theorems. A rich vein of ideas begins with Alexandrov's observation that, when considering a triangle Δ in a complete CAT(0) space X, if one gets any non-trivial equality in the CAT(0) inequality, then Δ spans an isometrically embedded triangular Euclidean disc in X. This observation leads one quickly to the fact that any pair of geodesic lines $\mathbb{R} \to X$ a bounded distance apart must bound a flat strip in X, and thence to a Product Decomposition Theorem:

Let $c\colon \mathbb{R} \to X$ be a geodesic line and let P be the set of geodesic lines c' contained in a bounded neighborhood of $c(\mathbb{R})$. Let $c'_0 \in c'(\mathbb{R})$ be the unique point closest to $c(0)$ and let $X^0_c = \{c'_0 \mid c' \in P\}$. Then $\bigcup\{c'(\mathbb{R}) \mid c' \in P\}$ is isometric to $X^0_c \times \mathbb{R}$.

This places severe restrictions on the way groups can act on CAT(0) spaces.

Proposition 2.1. *If Γ acts properly and cocompactly by isometries on a* CAT(0) *space and $\gamma \in \Gamma$ has infinite order, then the centralizer $C_\Gamma(\gamma)$ has a subgroup of finite index that splits as a direct product $C_0 = N \times \langle\gamma\rangle$.*

To prove this, one considers the union $\mathrm{Min}(\gamma)$ of the geodesic lines in X that are left invariant by γ. The Product Decomposition Theorem gives a splitting $\mathrm{Min}(\gamma) = \mathrm{Y}_0 \times \mathbb{R}$. The centralizer $C_\Gamma(\gamma)$ preserves $\mathrm{Min}(\gamma)$ and its splitting, acting by translations on the second factor. $C_\Gamma(\gamma)$ is finitely generated (3.3), so the image of $C_\Gamma(\gamma) \to \mathrm{Isom}(\mathbb{R})$ is isomorphic to $\mathbb{Z}^r$ for some r. Projecting onto a direct factor of $\mathbb{Z}^r$ gives an epimorphism ϕ from a subgroup of finite index $C_0 \subset C_\Gamma(\gamma)$ to $\langle\gamma\rangle$. Hence $C_0 = \ker\phi \times \langle\gamma\rangle$. A similar argument proves:

If a finitely generated group Γ acts by isometries on a CAT(0) *space* (*the action need not be proper*) *and a central subgroup $A \cong \mathbb{Z}^n$ acts freely by hyperbolic isometries, then Γ has a subgroup of finite index that contains A as a direct factor.*

These results give us our first glimpse of the fact that centralizers play an important role in non-positively groups. Many groups that lie in SH and Auto are seen not to lie in $\mathcal{C}_0$ because their centralizers do not virtually split; examples include mapping class groups [77] and central extensions of hyperbolic groups [79]. Such central extensions are defined by bounded cocycles and hence are quasi-isometric to the corresponding direct products. It follows that $\mathcal{C}_0$ is not closed under quasi-isometry.

More elaborate arguments akin to the one sketched above allow one to generalize splitting theorems proved in the Riemannian setting by D. Gromoll and J. Wolf [61] and B. Lawson and S. Yau [74] (see [35], pp. 239-253).

Theorem 2.2. *If a group $\Gamma = \Gamma_1 \times \Gamma_2$ with trivial centre acts properly and cocompactly by isometries on a* CAT(0) *space in which geodesics can be extended locally, then Γ_1 and Γ_2 also admit such actions.*

Theorem 2.3. *If Γ acts properly and cocompactly by isometries on a* CAT(0) *space X, then every solvable subgroup $S \subset \Gamma$ is finitely generated and virtually abelian. Moreover, S leaves invariant an isometrically embedded copy of Euclidean space $\mathbb{E}^n \hookrightarrow X$ on which is acts cocompactly.*

A refinement of the last part of this theorem can be used to identify constraints on the *length functions* $\gamma \mapsto \min_x d(x, \gamma.x)$ associated to semisimple actions on CAT(0) spaces. Such constraints serve as obstructions to the existence of actions both absolutely and in certain dimensions [18], [25], [57].

2.3. Subgroups in $\mathcal{C}_0$. As one pursues an understanding of the groups that act properly and cocompactly by isometries on CAT(0) spaces one finds increasingly subtle obstructions to the existence of metrics of non-positive curvature on aspherical spaces. In order to get at the heart of this subtlety one wants to bypass the obstructions to semisimple actions. One way of doing this is to focus on the finitely presented subgroups of fundamental groups of compact non-positively curved spaces. It transpires that such subgroups form a much more diverse class than the fundamental groups themselves – see [35], III.Γ.5, [29] and Section 5.

In very low dimensions, finitely presented subgroups are well-behaved [29]. *If Γ is the fundamental group of a compact non-positively curved manifold of dimension ≤ 3 (allowing boundary) then so too is each of its finitely generated subgroups.*

A similar result holds for complexes of dimension ≤ 2, except that one has to impose the hypothesis that the subgroups are finitely presented. In higher dimensions all manner of additional obstructions emerge: higher finiteness conditions, the complexity of decision problems, the structure of centralizers, *etc.* The subtlety of the situation is illustrated by the following construction [29].

Theorem 2.4. *There exist pairs of closed aspherical manifolds $N^n \hookrightarrow M^{n+1}$ with the following properties: M supports a metric of non-positive sectional curvature; $\pi_1 N \hookrightarrow \pi_1 M$ is a quasi-isometric embedding; the centralizers of all finite subsets in $\pi_1 N$ are fundamental groups of closed aspherical manifolds and have solvable word and conjugacy problems; but $\pi_1 N$ is not semihyperbolic, and hence N does not support a metric of non-positive curvature.*

One sees that $\pi_1 N$ is not semihyperbolic by examining the complexity of the word problem in centralizers: although $\pi_1 N$ satisfies a polynomial isoperimetric inequality, the centralizers of certain elements do not.

2.4. Complexes of group. An important instance of the local-to-global effect of non-positive curvature is the *Developability Theorem* for non-positively curved complexes of groups. This was inspired by Gromov and proved by Haefliger, who placed it in the more general setting of groupoids of local isometries [35], p. 584.

Complexes of groups were introduced by Haefliger to describe groups actions on 1-connected polyhedral complexes in terms of suitable local data on the quotient. If a complex of groups arises from such an action then it is said to be *developable*. In contrast to the 1-dimensional situation (graphs of groups), complexes of groups are not developable in general. But, crucially, they are if they satisfy a (local) *non-positive curvature condition* [35]. A full account of the theory is given in the final chapters of our book [35]. In recent years, this theory has played an important role in the construction of group actions.

3. Non-positively curved groups

In this section I'll describe the manifestations of non-positive curvature in group theory that arise from the following strategy. One starts by identifying a *robust feature* of CAT(0) spaces that encapsulates much of their large-scale geometry. Then, given a group Γ with generators $\mathcal{A}$ acting properly and cocompactly by isometries on a CAT(0) space X with basepoint p, one tries to articulate what remains of this feature when it is pulled-back to the Cayley graph $\mathcal{C}_{\mathcal{A}}(\Gamma)$ via the Γ-equivariant quasi-isometry sending the edge $[1, a]$ ($a \in \mathcal{A}$) to the geodesic $[p, a.p]$. One wants to define a group to be non-positively curved if it satisfies the resulting condition. The condition should be strong enough to facilitate a range of theorems analogous to what one knows about the prototypical groups of isometries, but one wants to avoid unnecessary hypotheses.

3.1. Hyperbolic groups. Let me recall how such a strategy is implemented in the hyperbolic case. The prototypical hyperbolic group is a group Γ that acts properly and cocompactly by isometries on a CAT(-1) space X. By comparing geodesic triangles $\Delta = \Delta(x, y, z)$ in X to triangles $\overline{\Delta} \subset \mathbb{H}^2$, one sees that there is a universal constant δ such that the distance from any point $q \in [x, y]$ to $[x, z] \cup [z, y]$ is at most δ. Moreover, quasigeodesics in CAT(-1) spaces stay uniformly close to geodesics, so (λ, ε)-quasigeodesic triangles in X are uniformly thin in the same sense (with a different δ). The quasi-isometry $\mathcal{C}_{\mathcal{A}}(\Gamma) \to X$ sends geodesic triangles in $\mathcal{C}_{\mathcal{A}}(\Gamma)$ to (λ, ε)-quasigeodesic triangles in X, where λ and ε depend on $\mathcal{A}$ and p. Therefore geodesic triangles in $\mathcal{C}_{\mathcal{A}}(\Gamma)$ are also uniformly thin. One takes this to be the defining property of a hyperbolic group.

Gromov's great insight is that because the thin triangles condition (which has many reformulations [35], p. 407) encapsulates so much of the essence of the large-scale geometry of CAT(-1) spaces, the groups whose Cayley graphs satisfy this condition share almost all of the properties enjoyed by the groups of isometries that were their prototypes. For example, every hyperbolic group Γ acts properly and cocompactly

on a contractible cell complex, has only finitely many conjugacy classes of finite subgroups, and contains no copy of $\mathbb{Z}^2$. Hyperbolic groups also enjoy a great deal of algorithmic structure. They are precisely the groups with linear Dehn functions. Their conjugacy problems can be solved in less than quadratic time, and conjugacy for finite subsets can also be determined efficiently [36]. Strikingly, the *translation lengths* $\tau(\gamma) = \lim d(1, \gamma^n)/n$ of elements of infinite order are rational numbers with bounded denominators [65], [51]. And given a finite generating set $\mathcal{A}$, the set of geodesic words for Γ is a *regular language*, i.e. there is a finite state automaton that recognises which words label geodesics in $\mathcal{C}_{\mathcal{A}}(\Gamma)$.

3.2. The pantheon of non-positively curved groups. The behaviour of geodesics described in (2.1) explains much of the global geometry of CAT(0) spaces, so we apply our strategy to this condition. For the prototype of Γ acting on X, the geodesics $[p, \gamma.p]$ pull-back to quasi-geodesics σ_γ connecting 1 to γ in $\mathcal{C}_{\mathcal{A}}(\Gamma)$. There is no loss of generality in assuming that σ_γ is an edge-path. By identifying σ_γ with the word in the generators $\mathcal{A}^{\pm 1}$ that labels it, we get a map $\sigma : \gamma \mapsto \sigma_\gamma$ to the free monoid $(\mathcal{A} \cup \mathcal{A}^{-1})^*$, where the image $\sigma(\Gamma)$ can be studied as a *formal language*. Following Bill Thurston [55], one calls σ a *combing*.

The convexity of the metric on X (2.1) implies that there exists a constant $k > 0$ such that for all $\gamma, \gamma' \in \Gamma$,

$$d(\sigma_\gamma(t),\ \sigma_{\gamma'}(t)) \leq k\, d(\gamma, \gamma') \tag{3.1}$$

for all $t \leq \max\{|\sigma_\gamma|,\ |\sigma_{\gamma'}|\}$. This is called the *fellow-traveller property*. A group that admits a combing (normal form) with this property is said to be *combable*.

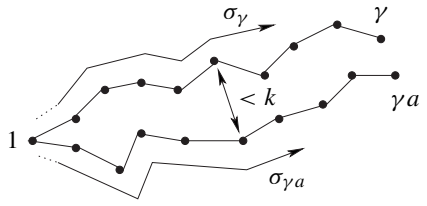


Figure 2. The fellow-traveller property.

All hyperbolic groups Γ are combable: one can take σ_γ to be any geodesic from 1 to γ but it is better to be systematic, choosing the first word in the dictionary-ordering obtained by ordering the generating set. If one adopts this systematic choice, then $\sigma(\Gamma)$ will be a *regular language*. With this in mind, ought one to include regularity as part of the definition of a non-positively curved group?

Requiring $\sigma(\Gamma)$ to be regular leads to the theory of *automatic groups*, which sprang from conversations between Bill Thurston and Jim Cannon on the algorithmic properties of Kleinian groups [46] and grew into a rich theory with large classes of

natural examples. It is described in detail in the book by Epstein *et al.* [55]. Automatic groups lend themselves well to practical computation.

If $\sigma(\Gamma)$ is a regular language, it follows from the fellow-traveller property that the paths σ_γ are quasi-geodesics with uniform constants. But in the absence of regularity it is unclear if one eliminates groups by imposing conditions on the length of σ_γ. This question is related to the complexity of the word problem for combable groups: a diagrammatic argument [21] shows that if one has a function $L\colon \mathbb{N} \to \mathbb{N}$ bounding the length of combing lines, $|\sigma_\gamma| \leq L(d(1,\gamma))$, then the Dehn function of Γ satisfies $\delta_\Gamma(n) \preccurlyeq n\,L(n)$; in particular, if the combing lines are quasi-geodesics then the group satisfies a quadratic isoperimetric inequality.

Another dilemma arises from the observation that although the convexity of $t \mapsto d(c_1(t), c_2(t))$ in the CAT(0) setting follows easily from the special case $c_1(0) = c_2(0)$, the analogous statement for groups is false. Thus it is unclear whether every group Γ that admits a combing with the fellow-traveller property must admit a combing σ with the stronger property

$$d(a.\sigma_{a^{-1}\gamma a'}(t),\ \sigma_\gamma(t)) \leq k \tag{3.2}$$

for all $a, a' \in \mathcal{A}$ and $\gamma \in \Gamma$. Groups that admit such combings are said to be *bicombable* [90]. If, in addition, the combing lines are quasi-geodesics (with uniform constants) then, following Alonso–Bridson [2], one says the group is *semihyperbolic*; this is the smallest of the classes we are discussing that includes $\mathcal{C}_0$.

This completes my brief sketch of how the classes Comb, Auto and SH marked in figure 1 present themselves for study. But are these classes distinct? Geometric group theory in the 1990s was marred by the absence of examples to distinguish between them but this situation was now been resolved [26].

Theorem 3.1. *There exist combable groups that are neither bicombable nor automatic.*

Once one knows that combable groups need not be automatic, it is natural to ask what classes or groups are incorporated if one places weaker constraints on the linguistic complexity of the formal language $\sigma(\Gamma) \subseteq (\mathcal{A} \cup \mathcal{A}^{-1})^*$. Among full abstract families of languages the regular, context-free and indexed languages form a hierarchy Reg $\subset$ CF $\subset$ Ind. If $\sigma(\Gamma)$ lies in a family $\mathcal{F}$, one calls σ an $\mathcal{F}$-combing.

Theorem 3.2 ([26]). *There exist* Ind*-combable groups that are not automatic; some of these have quadratic Dehn functions, others have cubic ones.*

A fascinating aspect of the struggle to understand different manifestations of non-positive curvature concerns the *free-by-cyclic groups* $F_n \rtimes \mathbb{Z}$. One feels these groups ought to conform to the expectations of non-positive curvature, yet they remain enigmatic. Since free-group automorphisms are more complicated than surface automorphisms, free-by-cyclic groups are more complicated than the fundamental groups $\Sigma_g \times \mathbb{Z}$ of 3-manifolds that fibre over the circle. We know that $F_n \rtimes_\phi \mathbb{Z}$ need not be

automatic [17] but we do not know for which ϕ it is; nor when it lies in $\mathcal{C}_0$. Daniel Groves and I used the train-track technology of [10], [11] to prove that all $F_n \rtimes \mathbb{Z}$ lie in IP(2), but this is highly non-trivial.

3.3. Subgroups. We saw in (2.2) that centralizers of groups in SH do not virtually split as they do in $\mathcal{C}_0$. However, SH is closed under passage to centralizers of finite subsets. Theorem 2.2 extends to SH. As for Theorem 2.3, one can prove that any polycyclic subgroup P of $\Gamma \in \mathrm{SH}$ must be virtually abelian and that $P \hookrightarrow \Gamma$ must be a quasi-isometric embedding, but it is unknown if abelian subgroups must be finitely generated. The positive part of this statement ultimately derives from the fact that translation numbers $\tau(\gamma) = \lim d(1, \gamma^n)/n$ are positive for elements of infinite order, while the negative part derives from the fact that, unlike in Hyp and $\mathcal{C}_0$, one does not know if the set of these numbers is discrete. These results are proved in [2] and [90] using ideas from [59].

In combable groups, the control over centralizers is lost [26].

3.4. The conjugacy and isomorphism problems. If two rectifiable loops c_0, c_1 in a compact, non-positively curved space X are freely homotopic, they are homotopic through loops c_t of length $l(c_t) \leq \max\{l(c_0),\ l(c_1)\}$. Accordingly, there is a constant $k > 0$ so that any conjugacy between words u_0, u_1 in the generators of $\pi_1 X$ can be realized by a sequence of moves $u_t \mapsto a_t u_t a_t^{-1}$ where a_t is a generator and $d(1,\ u_t) \leq K \max\{|u_0|,\ |u_1|\}$; see [35], p. 445. This control carries over to SH and bicombable groups, where it yields a solution to the conjugacy problem. But control is lost as one weakens the link to CAT(0) spaces [27].

Theorem 3.3. *The conjugacy problem is unsolvable in certain combable groups.*

Zlil Sela [88] solved the isomorphism problem for torsion-free hyperbolic groups. His solution was recently extended to a large class of relatively hyperbolic groups by F. Dahmani and D. Groves [48]. When combined with the topological rigidity theorem of T. Farrell and L. Jones [56], Sela's result implies that the homeomorphism problem is solvable among closed n-manifolds, $n \geq 5$, that admit a metric of negative curvature. The results of Farrell and Jones remain valid for non-positively curved manifolds and there is therefore considerable interest in the isomorphism problem in SH, $\mathcal{C}_0$ and the subclass consisting of the fundamental groups of such manifolds. These problems remain open, but beyond SH decidability is lost:

Theorem 3.4 ([27]). *The isomorphism problem is unsolvable among combable groups.*

To prove this one seeks a recursive sequence $\langle A \mid R_n \rangle$ of presentations of combable groups such that there is no algorithm that decides which are isomorphic. Starting with a hyperbolic group H in which there is no algorithm to decide when maps $\phi_n \colon F_r \to H$ are epic, one extends ϕ_n to a homomorphism $\hat{\phi}_n \colon F_{2r} \to H$ and defines $\langle A \mid R_n \rangle$ to be a certain presentation of $\Gamma *_{\Sigma(n)} \Gamma$, where $\Gamma = (\mathbb{Z}_2 * H) \times F_{2r}$ and $\Sigma(n) = \{(\hat{\phi}_n(x), x) \mid x \in F_{2r}\}$ is quasi-isometrically embedded.

3.5. Asynchronous combings and 3-manifolds. As one moves further from the CAT(0) setting and weakens the convexity condition on the metric, combings arise that only satisfy a weakened form of the fellow-traveller property (3.1): the paths σ_γ remain close only after *monotone reparameterization*.

Little of the strength of non-positive curvature remains in this definition but it does embrace a much larger class of groups, e.g. [22], [55]. Moreover, the amount of convexity retained is enough to provide a reasonable solution to the word problem and to ensure that these groups have classifying spaces with only finitely many cells in each dimension. Epstein *et al.* examined what happens when one requires $\sigma(\Gamma)$ to be regular. Bob Gilman [32] and I explored larger families of languages.

Epstein and Thurston [55] proved that the fundamental group of a compact 3-manifold M is automatic if and only if it satisfies a quadratic isoperimetric inequality (which excludes connected summands that are torus bundles over the circle with infinite holonomy). Gilman and I, building on [22], sharpened the negative part of their theorem and proved that by using indexed languages one can construct combings that encode the coarse geometry of any cocompact 3-manifold. (This result relies on the fact that 3-manifolds are geometrizable.)

Theorem 3.5. *The fundamental group of every compact* 3*-manifold M is asynchronously* Ind*-combable, but in some cases $\pi_1 M$ is not asynchronously* CF*-combable.*

4. Word problems and filling invariants

4.1. Dehn functions. The word problem for finitely presented groups has been at the heart of combinatorial group theory since its inception. When one attacks the word problem for a finitely presented group $\Gamma = \langle \mathcal{A} \mid \mathcal{R} \rangle$ directly, one's chances of success depend heavily on the *Dehn function* $\delta_\Gamma \colon \mathbb{N} \to \mathbb{N}$. Given a word w in the kernel of the map from the free group $F(\mathcal{A})$ to Γ, one defines

$$\mathrm{Area}(w) := \min\left\{N \mid w \overset{\text{free}}{=} \prod_{i=1}^{N} x_i^{-1} r_i x_i \text{ some } x_i \in F(\mathcal{A}),\ r_i \in \mathcal{R}^{\pm 1}\right\} \qquad (4.1)$$

and

$$\delta_\Gamma(n) := \max\{\mathrm{Area}(w) \mid w =_\Gamma 1,\ |w| \le n\}.$$

The subscript on δ_Γ is somewhat misleading since different finite presentations of the same group will in general yield different Dehn functions. This ambiguity is tolerated because it is tightly controlled: if the groups defined by two finite presentations are isomorphic, or just quasi-isometric, the corresponding Dehn functions are $\simeq$ equivalent in the following sense: *given two monotone functions $f, g \colon [0,\infty) \to [0,\infty)$, one writes $f \preccurlyeq g$ if there exists a constant $C > 0$ such that $f(l) \le C\, g(Cl + C) + Cl + C$ for all $l \ge 0$, and $f \simeq g$ if $f \preccurlyeq g$ and $g \preccurlyeq f$; and one extends this relations to include functions $\mathbb{N} \to [0,\infty)$.*

If $\delta_\Gamma(n) \preccurlyeq n^d$, one says that Γ satisfies a *polynomial isoperimetric inequality* of degree d. See [30] for references and basic facts about Dehn functions.

The first step from word problems to filling problems is provided by *van Kampen's Lemma* [70], which states that $\mathrm{Area}(w)$ is equal to the least number of 2-cells in any *van Kampen diagram* for w. Such a diagram describes a combinatorial *filling* (i.e. null-homotopy) for the loop labelled w in the 1-skeleton of the 2-complex $K = K(\mathcal{A}, \mathcal{R})$ described in the first paragraph of the introduction.

Suppose $\Gamma = \langle \mathcal{A} \mid \mathcal{R} \rangle$ acts properly and cocompactly by isometries on a Riemannian manifold X. Fix $p \in X$. If X is simply connected, the quasi-isometry $\gamma \mapsto \gamma.p$ extends to a Γ-equivariant map $\phi\colon \widetilde{K} \to X$. An edge-loop σ in $\mathcal{C}_{\mathcal{A}}(\Gamma) = \widetilde{K}^{(1)}$ defines a piecewise-geodesic loop $\phi \circ \sigma$ in X, and a van Kampen diagram that describes a filling of σ defines a singular disc in X with boundary $\phi \circ \sigma$. Conversely, any rectifiable loop c in X can be approximated by a word-like loop $\phi \circ \sigma$ whose length is linearly bounded by that of c and (more delicately) disc-fillings $g\colon \mathbb{D}^2 \to X$ with $g|_{\partial D} = c$ give rise to van Kampen diagrams for σ.

This line of thought, initiated by Gromov, suggests that the large-scale behaviour of Riemannian filling-discs – quantified by features such as area, radius, diameter *etc.* – should be translated to the setting of van Kampen diagrams. Then, in the spirit of van Kampen's Lemma, these features can be used to measure the complexity of the word problem in Γ. The following implementation of this strategy was described by Gromov and presented in detail in [30] (also [45]).

Let M be a compact Riemannian manifold with universal cover $\widetilde{M}$. Define the filling area $\mathrm{FArea}(c)$ of a rectifiable loop $c\colon S^1 \to \widetilde{M}$ to be the infimum of the areas of all Lipschitz maps $g\colon \mathbb{D} \to X$ where $\mathbb{D}$ is the standard 2-disc and $g|_{\partial D}$ is a monotone reparameterization of c. Consider

$$\mathrm{Fill}_M(l) := \sup\{\, \mathrm{FArea}(c) \mid c\colon S^1 \to \widetilde{M},\ \mathrm{length}(c) \leq l\},$$

the *genus zero, 2-dimensional, isoperimetric function* of M.

Theorem 4.1 (Filling Theorem). $\mathrm{Fill}_M(l) \simeq \delta_{\pi_1 M}(l)$.

A similar statement holds for isoperimetric functions of more general compact spaces with upper curvature bounds. Similar theorems also hold with *area* replaced by other invariants of the geometry of filling-discs. Among these, the most actively studied is intrinsic diameter (i.e. diameter measured in the induced length metric on the disc); in this case, FArea and the Dehn function are replaced by (intrinsic) *isodiametric functions*. When translated into algebra, bounds on intrinsic diameter correspond to bounds on the length of the conjugating words x_i in (4.1).

Results giving lower bounds on intrinsic diameter often proceed via extrinsic diameter, i.e. diameter measured in the metric on the ambient space. It was only recently that Tim Riley and I constructed the first compact manifolds for which the isodiametric functions corresponding to the choice intrinsic-versus-extrinsic have distinct asymptotic behaviour [43]. This extends a considerable body of work relating different aspects of the geometry of filling discs [64], [84].

4.2. The isoperimetric spectrum. A major theme in geometric group theory in the 1990s and into this century has been the struggle to determine which $\simeq$ classes of functions arise as Dehn functions. (I shall say little about the complementary challenge of calculating the Dehn functions of groups of special interest.)

The development of our knowledge can be charted by how the set

$$\mathrm{IP} = \{\rho \in [1,\infty) \mid f(n) = n^\rho \text{ is } \simeq \text{a Dehn function}\}$$

came to be understood. This set is called the *isoperimetric spectrum*. I should emphasize that it is far from the case that all Dehn functions are of the form n^α: there are non-polynomial Dehn functions such as $n^\alpha \log n$, as well as examples of small presentations with huge Dehn functions, e.g. faster than any iterated exponential (see 4.4). If Γ has unsolvable word problem, $\delta_\Gamma(n)$ will grow faster than any recursive function (indeed this serves as a definition of such groups).

The class of groups with linear Dehn functions coincides with the class of hyperbolic groups. The non-hyperbolic groups in $\mathcal{C}_0$ and Auto have quadratic Dehn functions. Certain combable groups have cubic Dehn functions (3.2), as does the 3-dimensional Heisenberg group. In about 1992, sequences of groups $(\Gamma_d)_{d\in\mathbb{N}}$ such that the Dehn function of Γ_d is polynomial of degree d were discovered by Gromov [64], Baumslag–Miller–Short [7], and Bridson–Pittet [42]. The literature now contains such sequences with all manner of additional properties. An example of a group whose Dehn function is polynomial of degree $d+1$ is $\mathbb{Z}^d \rtimes_\phi \mathbb{Z}$, where $\phi \in \mathrm{GL}(d,\mathbb{Z})$ has 1's on the diagonal and superdiagonal and zeroes elsewhere.

A result of Gromov [65], reproved by many people, states that if the Dehn function of a group is sub-quadratic (i.e. $\delta_\Gamma(n) = o(n^2)$) then $\delta_\Gamma(n) \simeq n$. Thus $\mathrm{IP} \cap (1,2)$ is empty. This begs the question of what other gaps there may be in IP, and whether there are any non-integral exponents at all. I settled this last question by constructing the *abc groups* of [20], formed by taking three torus bundles over the circle (different dimensions) and amalgamating their fundamental groups along central cyclic subgroups. Indiscrete families of exponents were first constructed in [86]. Noel Brady and I [15] completed the understanding of the coarse structure of IP by constructing a dense set of exponents in $[2,\infty)$.

Theorem 4.2. *The closure of* IP *is* $\{1\} \cup [2,\infty)$.

We proved this by associating to each pair of positive integers $p \geq q$ an aspherical 2-complex whose fundamental group

$$G_{p,q} \;=\; \langle a,b,s,t \mid [a,b]=1,\ sa^q s^{-1} = a^p b,\ ta^q t^{-1} = a^p b^{-1}\rangle,$$

has Dehn function $\simeq n^{2\log_2(2p/q)}$. These complexes are obtained by attaching a pair of annuli to a torus in a manner that ensures the existence of a family of discs in the universal cover that display a certain *snowflake geometry*. With Max Forester and Ravi Shankar [16], we developed a more sophisticated version of the snowflake

construction that yields a much larger class of isoperimetric exponents, showing in particular that $[2, \infty) \cap \mathbb{Q} \subseteq \mathrm{IP}$.

Once one knows that IP is not a discrete set, one assumes that it will follow the general pattern of group theory by exhibiting all plausible levels of complexity. This expectation is realized in a remarkable piece of work by M. Sapir, J.-C. Birget and E. Rips [86] who give a comprehensive description of Dehn functions $\delta_\Gamma(n) \succcurlyeq n^4$ by encoding the time functions of Turing machines. (The fine structure of $\mathrm{IP} \cap (2, 4)$ has yet to be determined.) In a subsequent work with A. Yu. Ol'shanskii [19] the same authors prove that the word problem for Γ is in NP if and only if Γ is a subgroup of a finitely presented group with polynomial Dehn function.

4.3. Groups with quadratic Dehn functions. The structure of IP provides us with two classes of groups that demand special attention – the groups with linear Dehn functions (which we know to be the hyperbolic groups) and the groups with quadratic Dehn functions. It is far from clear what to expect from groups in this second class. They have simply-connected asymptotic cones [80] but so do many (not all [23]) other groups with polynomial Dehn functions. It is unknown if they all have a solvable conjugacy problem. IP(2) contains many nilpotent groups N that are not virtually abelian and certain non-nilpotent polycyclic groups [53]. It is unknown if it contains any solvable groups that are not virtually polycyclic. Thurston proposed that $\mathrm{SL}(n, \mathbb{Z})$, $n \geq 4$, should be in IP(2) but this has yet to be confirmed. V. Guba [67] proved that Richard Thompson's group (which is torsion-free, of type FP_∞, and infinite dimensional) lies in IP(2). Groves and I proved the same for groups of the form $F_n \rtimes \mathbb{Z}$ [33].

4.4. Applications to the geometry and topology of manifolds. The dictionary of equivalence illustrated by the Filling Theorem translates information about Dehn functions into statements about the geometry of manifolds. But there are also less obvious mechanisms that allow one to gain geometric and topological information from an understanding of the nature of Dehn functions.

The *Andrews–Curtis conjecture* is one of the famous open problems of low-dimensional topology. It is related to the Zeeman conjecture and the smooth 4-dimensional Poincaré conjecture [72]. It asserts that one can reduce any *balanced presentation* $\langle a_1, \ldots, a_n \mid r_1, \ldots, r_n \rangle$ of the trivial group to the presentation $\langle a_1, \ldots, a_n \mid a_1, \ldots, a_n \rangle$ by a sequence of certain elementary moves. The main construction of [31] associates a balanced presentation P_w to each word w in the generators of a group B satisfying a deletion condition. P_w presents the trivial group if and only if $w = 1$ in B. Moreover, if P_w presents $\{1\}$ then it satisfies the Andrews–Curtis conjecture *but* the number of elementary moves required to trivialise it is bounded below by $\log \mathrm{Area}^*_B(w)$.

One gets dramatic lower bounds by taking $B = \langle a, t \mid [tat^{-1},\ a] = a^{r-1} \rangle$, since $\delta_B(n) \simeq \Delta_r \lfloor \log_2 n \rfloor$ where $\Delta_r(m)$ is defined by $\Delta_r(1) := r$ and $\Delta_r(m+1) := r^{\Delta_r(m)}$. In this case P_w has 4 generators and relations of total length $2(10 + |w| + r)$.

In a remarkable series of papers, A. Nabutovsky and S. Weinberger [78] use Dehn functions to explore the sub-level sets of functionals such as diameter on moduli spaces of metrics for closed manifolds M^n, $n \geq 5$. The constructions in [31] allow one to extend parts of their work to dimension 4.

4.5. Higher-dimensional isoperimetric inequalities. In the Riemannian context, having considered the isoperimetric problem for discs filling loops, it is natural to explore fillings of higher-dimensional spheres. In particular one wants to understand the isoperimetric function that bounds the volume of optimal ball-fillings. Correspondingly, one defines the *k-th order Dehn function* $\delta^{(k)}$ of a finitely presented group Γ that has a classifying space X with a compact $(k+1)$-skeleton $X^{(k+1)}$. Such functions were introduced by Gromov [64]. Roughly speaking $\delta^{(k)}(l)$ bounds the number of $(k+1)$-cells required to fill any singular k-sphere in $X^{(k)}$ comprised of at most l k-cells. The algebraic foundations of the subject were worked out carefully by Alonso *et al.* [3] and interpreted more topologically in [24]. From an algebraic point of view, $\delta^{(k)}(l)$ provides the least upper bound on the number of summands required to express an element $[f] \in \pi_k(X^{(k)})$ as a Γ-linear combination of the attaching maps of the $(k+1)$-cells of X. The $\simeq$ equivalence class of $\delta^{(k)}$ is an invariant of quasi-isometry.

In each dimension k one has the *isoperimetric spectrum*

$$\mathrm{IP}^{(k)} = \{\alpha \in [1,\infty) \mid f(x) = x^\alpha \text{ is } \simeq \text{ a } k\text{-th order Dehn function}\}.$$

Until recently, our knowledge even for $\mathrm{IP}^{(2)}$ was remarkably sparse, but my recent work with Brady, Forester and Shankar [16] remedies this. *We prove that if P is an irreducible non-negative integer matrix with Perron–Frobenius eigenvalue $\lambda > 1$, and r is an integer greater than every row sum in P, then for every $k \geq 2$ there is a group $\Gamma = \Sigma^{k-1} G_{r,P}$ with a compact $(k+1)$-dimensional classifying space such that $\delta^{(k)}(x) \simeq x^{2\log_\lambda(r)}$.* It follows from this and a related result in [16] that $\mathrm{IP}^{(k)}$ is dense in the range $[1+1/k, \infty)$. Indeed the case of 1×1 matrices alone leads to:

Theorem 4.3. $\mathbb{Q} \cap [1+\frac{1}{k}, \infty) \subset \mathrm{IP}^{(k)}$.

The exponent $1+1/k$ arises for $\mathbb{Z}^{k+1}$. Comparing with $\mathrm{IP} = \mathrm{IP}^{(1)}$, it is tempting to speculate that $\overline{\mathrm{IP}}^{(k)} = \{1\} \cup [1+1/k, \infty)$ but there are reasons to doubt this. One suspects that the fine structure of $\mathrm{IP}^{(k)}$ is similar to that of $\mathrm{IP}^{(1)}$.

The group $G_{r,P}$ is the fundamental group of an aspherical 2-complex $X_{r,P}$ assembled from a finite collection of annuli and tori; the rational number r encodes the multiplicities of the attaching maps while the matrix P encodes a prescription for the number and orientation of the tubes connecting each pair of tori. Least-area discs in $\widetilde{X}_{r,P}$ exhibit a more subtle form of the *snowflake geometry* from [15]. When r is an integer, certain families of these discs admit a precise scaling by a factor of r. One stacks scaled copies of them to form embedded 3-balls in the universal covering of the mapping torus associated to a certain 2-letter HNN extension $\Sigma G_{r,P}$ of $G_{r,P}$. These balls provide a lower bound on $\delta^{(2)}$ of $\Sigma G_{r,P}$; this proves to be sharp. The balls inherit the scaling property, so one can iterate.

The calculation of $\delta^{(k)}$ for $\Sigma^{(k-1)}G_{r,P}$ involves an induction on dimension. In order to make this work smoothly, one must bound not only the isoperimetric behaviour of disc-fillings for spheres but also the isoperimetric behaviour of *other pairs of compact manifolds* $(M, \partial M)$ mapping to the complexes in question. The topological approach to $\delta^{(k)}$ taken in [24] is well-adapted to such generalizations.

The *homological filling invariants* of the groups $\Sigma^{(k-1)}G_{r,P}$ exhibit a similar range of behaviour. Such invariants provide upper bounds on the size of (cellular) $(k+1)$-chains needed to fill k-cycles in the universal covering of a classifying space with finite $(k+1)$-skeleton; size is measured using the ℓ_1-norm associated to the cellular basis. These invariants are easier to work with than their homotopical counterparts and relate well to the Riemannian setting – see [64] and [55], Chapter 10.

5. Subdirect products of hyperbolic groups

The results in this section highlight a dichotomy in the behaviour of the finitely presented subgroups of direct products of hyperbolic groups: in general the structure of such subgroups can be fiendishly complicated; but for free groups and limit groups, these subgroups are remarkably controlled.

5.1. Encoding wildness. E. Rips [85] found a simple algorithm that associates to a finite presentation $\mathcal{Q}$ a short exact sequence $1 \to N \to H \to Q \to 1$, where Q is the group that $\mathcal{Q}$ presents, N is a 2-generator group, and H is a 2-dimensional hyperbolic group. To get H from $\mathcal{Q}$, one adds two new generators a_1, a_2, replaces the relations $r = 1$ of $\mathcal{Q}$ by relations $r = U_r(a_1, a_2)$ and adds a new relation $x_i^{-\varepsilon} a_j x_i^{\varepsilon} = V_{i,j,\varepsilon}(a_1, a_2)$ for each generator x_i in $\mathcal{Q}$ and $j = 1, 2$, $\varepsilon = \pm 1$; the words U_r and $V_{i,j,\varepsilon}$ are chosen to satisfy a small-cancellation condition. This construction depends on the specific presentation $\mathcal{Q}$, not just the group Q.

The flexibility of the Rips construction is such that (at the expense of increasing the number of generators of N) one can arrange for H to have additional properties such as being the fundamental group of a compact negatively curved 2-complex [93], [35], p. 225, or residually finite [94]. Thus one can encode all of the complexity of finite group-presentations (the lions of figure 1) into the finitely generated subgroups of such H. But such constructions say nothing about finitely presented subgroups because, by a theorem of R. Bieri [14], N is not finitely presentable if Q is infinite. The following theorem from [6] obviates this difficulty.

Theorem 5.1 (1-2-3 Theorem). *Suppose that* $1 \to N \to \Gamma \to Q \xrightarrow{p} 1$ *is exact. If* N *is finitely generated,* Γ *is finitely presented and* Q *is of type* F_3*, then the fibre-product* $P := \{(\gamma_1, \gamma_2) \mid p(\gamma_1) = p(\gamma_2)\} \subseteq \Gamma \times \Gamma$ *is finitely presented.*

The name of this theorem comes from the fact that the groups N, Γ, Q are assumed to be of type F_1, F_2 and F_3 respectively. (Recall that a group G is of type F_k if there exists a $K(G, 1)$ with compact k-skeleton.) The F_3 hypothesis says π_2 of a

presentation 2-complex for Q is finitely generated as a $\mathbb{Z}Q$-module. This allows one to control the relations among the generators of $N \times N$ (cf. [4]).

By combining the Rips construction and the 1-2-3 *Theorem, one can encode the complexities of arbitrary finitely presented groups directly into the structure of finitely presented subgroups of direct products of hyperbolic groups.*

An application of this principle is described it the next section. Several other applications are given in [6], one of which was refined in [40] to prove that *there exist* 2*-dimensional hyperbolic groups* Γ *such that there is no algorithm to decide isomorphism among the finitely presented subgroups of* $\Gamma \times \Gamma \times \Gamma$.

5.2. Subdirect products of surface groups. John Stallings [89] and Robert Bieri [13] showed that among the kernels of maps from direct products of free groups to abelian groups one finds a range of finiteness properties; in particular there exist finitely presented subgroups of $F_2 \times F_2 \times F_2$ whose third homology is not finitely generated and finitely presented subgroups S of a direct product of n free groups that are of type F_{n-1} with $H_n(S, \mathbb{Z})$ not finitely generated. Thus one senses that the wild behaviour observed above may continue among subdirect products of free groups, and indeed it does for finitely generated subgroups [76].

But Gilbert Baumslag and Jim Roseblade proved that the only finitely presented subgroups S of a direct product of *two* free groups are the obvious ones: such S are free or have a subgroup of finite index that is the product of its intersections with the factors. Howie, Miller, Short and I [39] discovered an analogous phenomenon in higher dimensions, cf. (5.3).

If a subgroup S *of a direct product of* n *free and surface groups is of type*[4] F_n *then* S *has a subgroup of finite index that is a direct product of free and surface groups.*

The case of surface groups is important because of its implications concerning *the fundamental groups of compact Kähler manifolds*. The work of Delzant and Gromov [52] shows that if such a group Γ is torsion-free and has sufficient multi-ended splittings, then there is an exact sequence $1 \to \mathbb{Z}^n \to \Gamma_0 \to S \to 1$, where S is a subdirect product of surface groups and $\Gamma_0 \subset \Gamma$ has finite index. Motivated by this, one would like to understand *all* finitely presented subdirect products of surface groups. In [41] Miller and I proved the following theorem and a weaker version (involving nilpotent quotients) for products of arbitrarily many surfaces.

Theorem 5.2. *If* S *is a finitely presented subgroup of a direct product of at most three surface groups, then either* S *is virtually a product of free and surface groups* (*the case where* S *is of type* F_3) *or else* S *is virtually the kernel of a map from a product of surface groups to an abelian group* (*the Stallings–Bieri situation*).

One hopes that a complete classification of the finitely presented subdirect products of free and surface groups may be within reach. What we have already proved shows

[4] It is enough that finite-index subgroups of S have $H_i(-, \mathbb{Z})$ finitely generated for $i \leq n$.

that *the conjugacy and membership problems are solvable for all finitely presented subgroups of direct products of surface groups.* This would not remain true if one were to replace surface groups by arbitrary 2-dimensional hyperbolic groups or Kleinian groups. Likewise, the splitting phenomenon for subgroups of type F_n does not extend to these classes. But it is does extend to limit groups.

5.3. Limit groups again. A finitely generated group L is *fully residually free* if for each finite subset $X \subset L$ there is a homomorphism to a finitely generated free group $\psi_X \colon L \to F$ that is injective on X. It is difficult to prove that such L are finitely presented but it is then easy to deduce that these are the limit groups defined in (1.2). The term *limit group* was coined by Sela to connote that these are the groups that occur as limits of stable sequences $\phi_n \colon G \to F$, where G is an arbitrary finitely generated group and *stable* means that for each $g \in G$ either $I_g = \{n \in \mathbb{N} : \phi_n(g) = 1\}$ or $J_g = \{n \in \mathbb{N} : \phi_n(g) \neq 1\}$ is finite; the *limit* is the quotient of G by $\{g \mid |I_g| = \infty\}$.

Such sequences arise when one studies $\mathrm{Hom}(G, F)$. A homomorphism $\phi \colon G \to F$ gives an action of G on the tree that is the Cayley graph of F, and it is profitable to examine sequences (ϕ_n) in the space of G-actions on $\mathbb{R}$-trees. By bringing to bear much of what is known about such spaces, Sela ([87] *et seq.*) obtains a finite parameterization of $\mathrm{Hom}(G, F)$ and a hierarchical decomposition of limit groups. His description of the *elementarily free groups* $\mathrm{EF} \subset \mathcal{L}$ solves a famous problem of A. Tarski. Similar results were obtained in a parallel project by O. Kharlampovich and A. Myasnikov [71] *et seq.* For an introduction to limit groups, see [10].

Jim Howie and I [37], [38] and H. Wilton [92] have been using Sela's work to explore the subgroup structure of limit groups and their direct products. The similarities with surface groups include:

Theorem 5.3 ([37]). *If $G_1, \ldots, G_n$ are elementarily free and $\Gamma \subset G_1 \times \cdots \times G_n$ is of type $\mathrm{FP_n}$, then there are finitely presented subgroups $H_i \subset G_i$ such that Γ is isomorphic to a finite-index subgroup of $H_1 \times \cdots \times H_n$.*

It is likely that this can be extended from EF to $\mathcal{L}$ as conjectured by Sela.

6. Two questions of Grothendieck

In this section I shall outline my solution with Fritz Grunewald to two problems concerning profinite completions and representations of groups that were posed by Alexander Grothendieck in 1970 [66]. The proof exemplifies two general points that I made earlier: the importance of being able to craft groups with specific properties, and the usefulness of the encodings into subdirect products.

6.1. Profinite completions. The profinite completion of a group Γ is the inverse limit of the directed system of finite quotients of Γ; it is denoted by $\hat{\Gamma}$. If Γ is residually finite, the natural map $\Gamma \to \hat{\Gamma}$ is injective. In [66] Grothendieck related the representation theory of a finitely generated group to its profinite completion:

if $A \neq 0$ is a commutative ring and $u\colon \Gamma_1 \to \Gamma_2$ is a homomorphism of finitely generated groups, then $\hat{u}\colon \hat{\Gamma}_1 \to \hat{\Gamma}_2$ is an isomorphism if and only the restriction functor $u_A^*\colon \mathrm{Rep}_A(\Gamma_2) \to \mathrm{Rep}_A(\Gamma_1)$ is an equivalence of categories, where $\mathrm{Rep}_A(\Gamma)$ is the category of finitely presented A-modules with a Γ-action.

Grothendieck investigated under what circumstances $\hat{u}\colon \hat{\Gamma}_1 \to \hat{\Gamma}_2$ being an isomorphism implies that u is an isomorphism. This led him to pose the following problem: *Let Γ_1 and Γ_2 be finitely presented, residually finite groups and let $u\colon \Gamma_1 \to \Gamma_2$ be a homomorphism such that $\hat{u}\colon \hat{\Gamma}_1 \to \hat{\Gamma}_2$ is an isomorphism of profinite groups. Does it follow that u is an isomorphism from Γ_1 onto Γ_2?*

A negative solution to the corresponding problem for finitely generated groups was given by Platonov and Tavgen [81]. But there is an emphasis on finite presentability in Grothendieck's problem because of his original motivation for studying profinite completions: he wanted to understand the extent to which the topological fundamental group of a complex projective variety determines the algebraic fundamental group, and *vice versa*. In the Spring of 2003 Fritz Grunewald and I settled Grothendieck's question in the negative.

Theorem 6.1. *There exist residually finite, 2-dimensional, hyperbolic groups H and finitely presented subgroups $P \subseteq \Gamma := H \times H$ of infinite index, such that P is not abstractly isomorphic to Γ, but $u\colon P \hookrightarrow \Gamma$ induces an isomorphism $\hat{u}\colon \hat{P} \to \hat{\Gamma}$.*

The first ingredient in the proof is the following distillation of arguments of Platonov and Tavgen [81]. *Let $1 \to N \to H \to Q \to 1$ be a short exact sequence of groups with fibre product $u\colon P \hookrightarrow H \times H$. If Q is superperfect*[5] *and has no finite quotients, and H is finitely generated, then $\hat{u}\colon \hat{P} \to \hat{H} \times \hat{H}$ is an isomorphism.*

One applies this criterion to the output of an algorithm obtained by combining D. Wise's refinement [94] of the Rips construction with the 1-2-3 Theorem [6]:

There is an algorithm that associates to any finite aspherical presentation $\mathcal{Q}$ a short exact sequence $1 \to N \to H \to Q \to 1$ and a finite presentation for the fibre product $P \subset H \times H$, where H is hyperbolic and residually finite.

To complete the proof, one needs suitable input presentations $\mathcal{Q}$. A simple calculation in homology shows that if a perfect group has a balanced presentation then it is superperfect. Thus it suffices to construct balanced, aspherical presentations of infinite groups with no non-trivial finite quotients. The following example was constructed in [34]; earlier examples are due to G. Higman [69]. Let $p \geq 2$.

$$\langle a_1, a_2, \hat{a}_1, \hat{a}_2 \mid a_1^{-1}a_2^p a_1 a_2^{-p-1},\ \hat{a}_1^{-1}\hat{a}_2^p \hat{a}_1 \hat{a}_2^{-p-1},\ a_1^{-1}[\hat{a}_2, \hat{a}_1^{-1}\hat{a}_2\hat{a}_1],\ \hat{a}_1^{-1}[a_2, a_1^{-1}a_2 a_1]\rangle.$$

At the expense of complicating the construction of the *Grothendieck pair* $P \hookrightarrow \Gamma \times \Gamma$, one can replace the requirement that the input presentation $\mathcal{Q}$ be aspherical by the

[5] $H_1(Q, \mathbb{Z}) = H_2(Q, \mathbb{Z}) = 0$.

hypothesis that Q be of type F_3. This allows one to associate a Grothendieck pair to any group of type F_3; for if $\mathcal{G}$ is a class (such as F_3) closed under the formation of HNN extensions and amalgamated free products along free groups, one can embed any group $G \in \mathcal{G}$ into a $\overline{G} \in \mathcal{G}$ that has no finite quotients [28].

6.2. Grothendieck's Tannaka duality groups. In the same paper [66] as he raised the problem described above, Grothendieck described an idea for reconstructing a residually finite group from the tensor product structure of its representation category $\mathrm{Rep}_A(\Gamma)$. He encoded this structure into a Tannaka duality group: if $\mathrm{Mod}(A)$ is the category of all finitely generated A-modules and $\mathcal{F}\colon \mathrm{Rep}_A(\Gamma) \to \mathrm{Mod}(A)$ is the forgetful functor, Grothendieck defines $\mathrm{cl}_A(\Gamma)$ to be the group of natural self-transformations of the functor $\mathcal{F}$ that are compatible with the tensor product $\otimes_A$. And he poses the following problem: *If Γ is a finitely presented, residually finite group, is the natural monomorphism from Γ to $\mathrm{cl}_A(\Gamma)$ an isomorphism for every non-zero commutative ring A, or at least some suitable commutative ring $A \neq 0$?*

Theorem 6.2 ([34]). *If P is one of the groups constructed in Theorem* 6.1*, then P is of infinite index in $\mathrm{cl}_A(P)$ for every commutative ring $A \neq 0$.*

Previously, Alex Lubotzky [75] had exhibited finitely presented, residually finite groups Γ such that $\Gamma \to \mathrm{cl}_A(\Gamma)$ is not surjective when $A = \mathbb{Z}$.

References

[1] Alibegović, E., Bestvina, M., Limit groups are CAT(0). *J. London Math. Soc.*, in press.

[2] Alonso, J., Bridson, M. R., Semihyperbolic groups. *Proc. London Math. Soc.* **70** (1995), 56–114.

[3] Alonso, J. M., Wang, X., Pride, S. J., Higher-dimensional isoperimetric (or Dehn) functions of groups. *J. Group Theory* **2** (1999), 81–112.

[4] Baik, Y. G., Harlander, J., Pride, S. J., The geometry of group extensions. *J. Group Theory* **1** (1998), 395–416.

[5] Ballmann, W., *Lectures on spaces of nonpositive curvature.* DMV Seminar 25, Birkhäuser, Basel 1995.

[6] Baumslag, G., Bridson, M. R., Miller, C. F., Short, H., Fibre products, non-positive curvature and decision problems. *Comm. Math. Helv.* **75** (2000), 457–477.

[7] Baumslag, G., Miller, C. F., Short, H., Isoperimetric inequalities and the homology of groups. *Invent. Math.* **113** (1993), 531–560.

[8] Baumslag, G., Roseblade, J. E., Subgroups of the direct product of two free groups. *J. London Math. Soc.* **30** (1984), 44–52.

[9] Bestvina, M., The topology of $\mathrm{Out}(F_n)$. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 373–384.

[10] Bestvina M., Feighn, M., Notes on Sela's work: Limit groups and Makanin-Razborov diagrams. To appear in *Geometric and cohomological methods in group theory* (ed. by M. R. Bridson, P. H. Kropholler, I. J. Leary).

[11] Bestvina, M., Feighn, M., Handel, M., The Tits alternative for Out(F_n) II: A Kolchin type theorem. *Ann. of Math.* **161** (2005), 1–59.

[12] Bestvina, M., Handel, M., Train tracks and automorphisms of free groups. *Ann. of Math.* **135** (1992), 1–51.

[13] Bieri, R., Normal subgroups in duality groups and in groups of cohomological dimension 2. *J. Pure Appl. Algebra* **7** (1976), 35–51.

[14] Bieri, R., *Homological dimension of discrete groups*. Queen Mary Math. Notes, London 1976.

[15] Brady, N., Bridson, M. R., There is only one gap in the isoperimetric spectrum. *Geom. Funct. Anal.* **10** (2000), 1053–1070.

[16] Brady, N., Bridson, M. R., Forester, M., Shankar, K., Snowflake groups, Perron-Frobenius exponents and isoperimetric spectra. Preprint, 2006.

[17] Brady, N., Bridson, M. R., Reeves, L., Free-by-cyclic groups that are not automatic. Preprint, 2006.

[18] Brady, N., Crisp, J., Two-dimensional Artin groups with CAT(0) dimension three. *Geom. Dedicata* **94** (2002), 185–214.

[19] Birget, J-C., Olshanskii, A.Yu., Rips, E., Sapir, M. V., Isoperimetric functions of groups and computational complexity of the word problem. *Ann. of Math.* **156** (2002), 467–518.

[20] Bridson, M. R., Fractional isoperimetric inequalities and subgroup distortion. *J. Amer. Math. Soc.* **12** (1999), 1103–1118.

[21] Bridson, M. R., On the geometry of normal forms in discrete groups. *Proc. London Math. Soc.* **67** (1993), 516–616.

[22] Bridson, M. R., Combing semidirect products and 3-manifold groups. *Geom. Funct. Anal.* **3** (1993), 263–278.

[23] Bridson M.R., Asymptotic cones and polynomial isoperimetric inequalities. *Topology* **38** (1999), 543–554.

[24] Bridson, M. R., Polynomial Dehn functions and the length of asynchronous automatic structures. *Proc. London Math. Soc.* **85** (2002), 441–466.

[25] Bridson, M. R., Length functions, non-positive curvature and the dimension of discrete groups. *Math. Res. Lett.* **8** (2001), 557–567.

[26] Bridson, M. R., Combings of groups and the grammar of reparameterisation. *Comment. Math. Helv.* **78** (2003), 752–771.

[27] Bridson, M. R., The conjugacy and isomorphism problems for combable groups. *Math. Ann.* **327** (2003), 305–314.

[28] Bridson, M. R., Controlled embeddings into groups that have no non-trivial finite quotients, *Geom. Topol. Monogr.* **1** (1998), 99–116.

[29] Bridson, M. R., On the subgroups of semihyperbolic groups. In *Essays on geometry and related topics* (ed. by E. Ghys et al.), Vol. 1, Monogr. Enseign. Math. 38, L'Enseignement Mathématique, Geneva 2001, 85–111.

[30] Bridson, M. R., The geometry of the word problem. In *Invitations to geometry and topology* (ed. by M. R. Bridson, S. M. Salamon), Oxford University Press, Oxford 2002, 29–91.

[31] Bridson, M. R., On the complexity of balanced presentations and the Andrews-Curtis conjecture. Preprint, 2006.

[32] Bridson, M. R., Gilman, R., Formal language theory and the geometry of 3-manifolds. *Comment. Math. Helv.* **71** (1996), 525–555.

[33] Bridson, M. R., Groves, D. P., The quadratic isoperimetric inequality for mapping tori of free group automorphisms I and II. ArXiv math.GR/0211459 and GR/0507589.

[34] Bridson, M. R., Grunewald, F., Grothendieck's problems concerning profinite completions and representations of groups. *Ann. of Math.* **160** (2004), 359–373.

[35] Bridson, M. R., Haefliger, A., *Metric spaces of non-positive curvature.* Grundlehren Math. Wiss. 319, Springer-Verlag, Heidelberg 1999.

[36] Bridson, M. R., Howie, J., Conjugacy of finite subsets in hyperbolic groups. *Intl. J. Alg. Comp.* **15** (2005), 725–756.

[37] Bridson, M. R., Howie, J., Subgroups of direct products of elementarily free groups. *Geom. Funct. Anal.*, in press.

[38] Bridson, M. R., Howie, J., Subgroups of direct products of two limit groups. ArXiv math.GR/0510353.

[39] Bridson, M. R., Howie J., Miller C. F., Short H., The subgroups of direct products of surface groups. *Geom. Dedicata* **92** (2002), 95–103.

[40] Bridson, M. R., Miller, C. F., Recognition of subgroups of direct products of hyperbolic groups. *Proc. Amer. Math. Soc.* **132** (2004), 59–65.

[41] Bridson, M. R., Miller, C. F., Structure and finiteness properties for subdirect products of groups, Preprint, 2006.

[42] Bridson, M. R., Pittet, C., Isoperimetric inequalities for the fundamental groups of torus bundles over the circle. *Geom. Dedicata* **49** (1994), 203–219.

[43] Bridson, M. R., Riley, T., Intrinsic versus extrinsic diameter for Riemannian filling discs and van Kampen diagrams. ArXiv math.GR/0511004.

[44] Bridson, M. R., Vogtmann, K., Automorphism groups of free groups, surface groups and free abelian groups. ArXiv math.GR/0507612.

[45] Burillo, J., Taback, J., Equivalence of geometric and combinatorial Dehn functions. *New York J. Math.* **8** (2002), 169–179.

[46] Cannon, J. W., The combinatorial structure of cocompact discrete hyperbolic groups. *Geom. Dedicata* **16** (1984), 123–148.

[47] Champetier C., Guirardel, V., Limit groups as limits of free groups: compactifying the set of free groups. *Israel J. Math.* **146** (2005), 1–76.

[48] Dahmani, F., Groves, D., The isomorphism problem for toral relatively hyperbolic groups. ArXiv math.GR/0512605.

[49] Davis, M. W., Nonpositive curvature and reflection groups. In *Handbook of geometric topology*, North-Holland, Amsterdam 2002, 373–422.

[50] Dehn, M., Über unendliche diskontinuierliche Gruppen. *Math. Ann.* **71** (1912), 116–144.

[51] Delzant, T., Sous-groupes distingués et quotients des groups hyperboliques. *Duke Math. J.* **83** (1996), 661–682.

[52] Delzant, T., Gromov, M., Cuts in Kähler groups. In *Infinite Groups: Geometric, Combinatorial and Dynamical Aspects*, Progr. Math. 248, Birkhäuser, Basel 2005, 31–55.

[53] Druţu, C., Filling in solvable groups and in lattices in semisimple groups. *Topology* **43** (2004), 983–1033.

[54] Druţu, C., Sapir, M.V., Tree-graded spaces and asymptotic cones of groups (with appendix by D. Osin and M. Sapir). *Topology* **44** (2005), 959–1058.

[55] Epstein, D. B. A., Cannon, J. W., Holt, D. F., Levy, S. V. F., Paterson, M. S., Thurston, W. P., *Word Processing in Groups*. A.K. Peters, Boston, MA, 1992.

[56] Farrell, F. T., Jones, L. E., Topological rigidity for compact non-positively curved manifolds. In *Differential geometry: Riemannian geometry* (Los Angeles, CA, 1990), Proc. Sympos. Pure Math. 54, Amer. Math. Soc., Providence, RI, 1993, 229–274

[57] Gersten, S. M., The automorphism group of a free group is not a CAT(0) group. *Proc. Amer. Math. Soc.* **121** (1994), 999–1002.

[58] Gersten, S. M., Dehn functions and l_1-norms for finite presentations. In *Algorithms and Classification in Combinatorial Group Theory* (ed. by G. Baumslag, C. F. Miller), Math. Sci. Res. Inst. Publ. 23, Springer-Verlag, New York 1992, 195–224.

[59] Gersten, S. M., Short, H. B., Rational subgroups of biautomatic groups. *Ann. of Math.* **134** (1991), 125–158.

[60] Grigorchuk, R. I., Degrees of growth of finitely generated groups and the theory of invariant means. *Izv. Akad. Nauk SSSR* Ser. Mat. **48** (1984), 939–985.

[61] Gromoll, D., Wolf, J., Some relations between the metric structure and the algebraic structure of the fundamental group in manifolds of non-positive curvature. *Bull. Amer. Math. Soc.* **77** (1971), 545–552.

[62] Gromov, M., Infinite groups as geometric objects. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 385–392.

[63] Gromov, M., Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[64] Gromov, M., Asymptotic invariants of infinite groups. *Geometric Group Theory* (ed. by G. A. Niblo, M. A. Roller), Vol. 2, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge 1993.

[65] Gromov, M., Hyperbolic groups. In *Essays in Group Theory* (ed. by S. M. Gersten), Springer-Verlag, New York 1987, 75–263.

[66] Grothendieck,, A., Représentations linéaires et compactification profinie des groupes discrets. *Manuscripta Math.* **2** (1970), 375–396.

[67] Guba, V. S., The Dehn function of Richard Thompson's group F is quadratic. *Invent. Math.* **163** (2006), 313–342.

[68] Higman, G., Subgroups of finitely presented groups. *Proc. Roy. Soc. Ser. A* **262** (1961), 455–475.

[69] Higman, G., A finitely generated infinite simple group, *J. London Math. Soc.* **26** (1951), 61–64.

[70] van Kampen, E. R., On some lemmas in the theory of groups. *Amer. J. Math.* **55** (1933), 268–273.

[71] Kharlampovich, O. G., Myasnikov, A. G., Irreducible affine varieties over a free group I; II. *J. Algebra* **200** (1998), 472–516; 517–570.

[72] Kirby, R., Problems in low-dimensional topology. In *Geometric topology* (Athens, GA, 1993), AMS/IP Stud. Adv. Math. 2.2, Amer. Math. Soc., Providence, RI, 1997, 35–473.

[73] Kleiner, B., Leeb, B., Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings. *Inst. Hautes Études Sci. Publ. Math.* **86** (1997), 115–197.

[74] Lawson, H. B., Yau, S. T., Compact manifolds of nonpositive curvature. *J. Differential Geom.* **7** (1972), 211–228.

[75] Lubotzky, A., Tannaka duality for discrete groups. *Amer. J. Math.* **102** (1980), 663–689.

[76] Miller, C. F., *On group-theoretic decision problems and their classification.* Ann. of Math. Stud. 68, Princeton University Press, Princeton, N.J., 1971.

[77] Mosher, L., Mapping class groups are automatic. *Ann. of Math.* **142** (1995), 303–384.

[78] Nabutovsky, A., Weinberger, S., The fractal nature of Riem/Diff. I. *Geom. Dedicata* **101** (2003), 1–54.

[79] Neumann, W., Reeves, L., Central extensions of word hyperbolic groups. *Ann. of Math.* **145** (1997), 183–192.

[80] Papasoglu, P., On the asymptotic cone of groups satisfying a quadratic isoperimetric inequality. *J. Differential Geom.* **44** (1996), 789–806.

[81] Platonov, V., Tavgen, O. I., Grothendieck's problem on profinite completions of groups. *Soviet Math. Dokl.* **33** (1986), 822–825.

[82] Remeslennikov, V. N., ∃-free groups. *Sibirsk. Mat. Zh.* **30** (1989), 193–197; English transl. *Siberian Math. J.* **30** (1989), 998–1001.

[83] Riley, T. R., Higher connectedness of asymptotic cones. *Topology* **42** (2003), 1289–1352.

[84] Riley, T. R., Filling functions. ArXiv math.GR/0603059.

[85] Rips, E., Subgroups of small cancellation groups. *Bull. London Math. Soc.* **14** (1982), 45–47.

[86] Sapir, M. V., Birget, J.-C., Rips, E., Isoperimetric and isodiametric functions of groups, *Ann. of Math.* **157** (2002), 345-466.

[87] Sela, Z., Diophantine geometry over groups. I. Makanin-Razborov diagrams. *Inst. Hautes Études Sci. Publ. Math.* **93** (2001), 31–105.

[88] Sela, Z., The isomorphism problem for hyperbolic groups I. *Ann. of Math.* **141** (1995), 217–283.

[89] Stallings, J. R., A finitely presented group whose 3–dimensional homology group is not finitely generated. *Amer. J. Math.* **85** (1963) 541–543.

[90] Short, H., Groups and combings. Preprint, ENS Lyon, 1990.

[91] Vogtmann, K., Automorphisms of free groups and Outer Spaces. *Geom. Dedicata* **94** (2002), 1–31.

[92] Wilton, H., Subgroup separability of limit groups. PhD thesis, Univ. London, 2006.

[93] Wise, D., Incoherent negatively curved groups. *Proc. Amer. Math. Soc.* **126** (1998), 957–964.

[94] Wise, D., A residually finite version of Rips's construction. *Bull. London Math. Soc.* **35** (2003), 23–29.

Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom
E-mail: m.bridson@imperial.ac.uk

Link homology and categorification

Mikhail Khovanov

Abstract. This is a short survey of algebro-combinatorial link homology theories which have the Jones polynomial and other link polynomials as their Euler characteristics.

Mathematics Subject Classification (2000). 57M25, 57Q45.

Keywords. Link homology, quantum link invariants, matrix factorizations.

1. Introduction

The discovery of the Jones polynomial by V. Jones [J] and quantum groups by V. Drinfeld and M. Jimbo led to an explosive development of quantum topology. The newly found topological invariants were christened "quantum invariants"; for knots and links they often take the form of polynomials. By late 80s to early 90s it was realized that each complex simple Lie algebra $\mathfrak{g}$ gives rise to a gaggle of quantum invariants. To a link L in $\mathbb{R}^3$ with each component colored by an irreducible representation of $\mathfrak{g}$ there is assigned an invariant $P(L, \mathfrak{g})$ taking values in the ring of Laurent polynomials $\mathbb{Z}[q, q^{-1}]$ (sometimes fractional powers of q are necessary). Polynomials $P(L, \mathfrak{g})$ have a representation-theoretical description, via intertwiners between tensor products of irreducible representations of the quantum group $U_q(\mathfrak{g})$, the latter a Hopf algebra deformation of the universal enveloping algebra of $\mathfrak{g}$. These invariants by no means exhaust all quantum invariants of knots and links; various modifications and generalizations include finite type (Vassiliev) invariants, invariants associated with quantum deformations of Lie superalgebras, etc.

Quantum $\mathfrak{sl}(n)$ link polynomials, when each component of L is colored by the fundamental n-dimensional representation, can be conveniently encapsulated into a single 2-variable polynomial $P(L)$, known as the HOMFLY or HOMFLY-PT polynomial [HOMFLY], [PT].

The skein relation

$$\lambda P(L_1) - \lambda^{-1} P(L_2) = (q - q^{-1}) P(L_3)$$

for any three links L_1, L_2, L_3 that differ as shown below and the value of P on the unknot, uniquely determines the HOMFLY-PT invariant, which lies in the ring $\mathbb{Z}[\lambda^{\pm 1}, (q - q^{-1})^{\pm 1}]$ (in the original papers a single variable was used instead of

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$q - q^{-1}$, making P a genuine Laurent polynomial in two variables).

L_1 L_2 L_3

Specializing $\lambda = q^n$, for $n \geq 0$, leads to a link polynomial invariant $P_n(L) \in \mathbb{Z}[q, q^{-1}]$, normalized so that $P_n(\text{unknot}) = q^{n-1} + q^{n-3} + \cdots + q^{1-n}$ for $n > 0$ and $P_0(\text{unknot}) = 1$.

$P_0(L)$ and $P_2(L)$ are the Alexander and Jones polynomials of L, respectively, while $P_1(L)$ is a trivial invariant. For $n > 0$, the polynomial $P_n(L)$ can be interpreted via the representation theory of quantum $\mathfrak{sl}(n)$, and $P_0(L)$ – via that of the quantum Lie superalgebra $U_q(\mathfrak{gl}(1|1))$.

The miracle that emerged in the past few years is that these polynomials are Euler characteristics of link homology theories:

- The Jones polynomial $P_2(L)$ is the Euler characteristic of a bigraded link homology theory $\mathcal{H}(L)$, discovered in [K1].
- The Alexander polynomial $P_0(L)$ is the Euler characteristic of a bigraded knot homology theory, discovered by P. Ozsváth, Z. Szabó [OS1] and J. Rasmussen [R1].
- The polynomial $P_3(L)$ is the Euler characteristic of a link homology theory $\mathrm{H}(L)$, defined in [K2].
- For each $n \geq 1$, Lev Rozansky and the author constructed a bigraded link homology theory $\mathrm{H}_n(L)$ with $P_n(L)$ as the Euler characteristic, see [KR1].
- The entire HOMFLY-PT polynomial is the Euler characteristic of a triply-graded link homology theory [KR2], [K6] (for a possible alternative approach via string theory see [GSV]).

Ideally, a link homology theory should be a monoidal functor $\mathcal{F}$ from the category LCob of link cobordisms to a tensor triangulated category T (for instance, T could be the category of complexes of R-modules, up to chain homotopies, for a commutative ring R). Objects of LCob are oriented links in $\mathbb{R}^3$, morphisms from L_0 to L_1 are isotopy classes (rel boundary) of oriented surfaces S smoothly and properly embedded in $\mathbb{R}^3 \times [0, 1]$ such that $L_0 \sqcup (-L_1)$ is the boundary of S and $L_i \subset \mathbb{R}^3 \times \{i\}, i = 0, 1$. In many known examples, $\mathcal{F}$ is a projective functor: the map $\mathcal{F}(S)\colon \mathcal{F}(L_0) \longrightarrow \mathcal{F}(L_1)$ is well defined up to overall multiplication by invertible central elements of T (e.g. by ± 1 for homology theory $\mathcal{H}$).

No a priori reason why quantum link invariants should lift to link homology theories is known and the general framework for lifting quantum invariants to homology

theories remains a mystery. We call such a lift a *categorification* of the invariant. The term categorification was coined by L. Crane and I. Frenkel [CF] in the context of lifting an n-dimensional TQFT to an $(n+1)$-dimensional one ($n = 2, 3$ are the main interesting cases).

Let us also point out that the Casson invariant (a degree two finite-type invariant of 3-manifolds) is the Euler characteristic of instanton Floer homology, that the Seiberg–Witten and Ozsváth–Szabó 3-manifold homology theories categorify degree one finite-type invariants of 3-manifolds (the order of $H_1(M, \mathbb{Z})$ when the first homology of the 3-manifold M is finite and, more generally, the Alexander polynomial of M), that equivariant knot signatures are Euler characteristics of $\mathbb{Z}/4\mathbb{Z}$-graded link homologies (O. Collin, B. Steer [CS], W. Li), and that there exist ideas on how to categorify the 2-variable Kauffman polynomial [GW], the colored Jones polynomial, and quantum invariants of links colored by arbitrary fundamental representations $\Lambda^i V$ of $\mathfrak{sl}(n)$ [KR1].

2. A categorification of the Jones polynomial

In the late nineties the author discovered a homology theory $\mathcal{H}(L)$ of links which is bigraded,

$$\mathcal{H}(L) = \bigoplus_{i,j\in\mathbb{Z}} \mathcal{H}^{i,j}(L),$$

and has the Jones polynomial as the Euler characteristic,

$$P_2(L) = \sum_{i,j\in\mathbb{Z}} (-1)^i q^j \operatorname{rk}(\mathcal{H}^{i,j}(L)).$$

The construction of $\mathcal{H}$ categorifies the Kauffman bracket description of the Jones polynomial. Starting from a plane projection D of L we build homology groups $\mathcal{H}(D)$ inductively on the number of crossings of the projection via long exact sequences

and then check that $\mathcal{H}(D)$ are invariants of L alone. Homology of the empty link is $\mathbb{Z}$, homology of the unknot is $\mathcal{A} = \mathbb{Z}[X]/(X^2)$, which should be thought of as the integral cohomology ring of the 2-sphere. Homology of the k-component unlink is $\mathcal{A}^{\otimes k}$. The obvious cobordisms between unlinks turn $\mathcal{A}$ into a commutative Frobenius ring, with the trace map $\operatorname{tr}(1) = 0$, $\operatorname{tr}(X) = 1$ (in any full-fledged link homology theory homology of the unknot is a commutative Frobenius algebra over homology of the empty link). $\mathcal{H}(D)$ is the homology of a complex $\mathcal{C}(D)$ constructed in an elementary way from direct sums of tensor powers of $\mathcal{A}$ and the structure maps of this Frobenius ring.

Theorem 2.1. *There exists a combinatorially defined bigraded homology theory $\mathcal{H}(L)$ of oriented links in $\mathbb{R}^3$. The groups $\mathcal{H}^{i,j}(L)$ are finitely-generated and their Euler characteristic is the Jones polynomial. The theory is functorial: to an oriented cobordism S between links L_0 and L_1 it assigns a homomorphism of groups*

$$\mathcal{H}(S)\colon \mathcal{H}(L_0) \longrightarrow \mathcal{H}(L_1),$$

well defined up to overall minus sign and of bidegree $(0, -\chi(S))$, where $\chi(S)$ is the Euler characteristic of the surface S.

That $\pm\mathcal{H}(S)$ is well defined was proved in [Ja] and [K4] in two different ways.

The homology theory $\mathcal{H}$ is manifestly combinatorial and programs computing it were written by D. Bar-Natan, A. Shumakovitch and J. Green. The earliest program [BN1] led to the conjecture that ranks of the homology groups of alternating links are determined by the Jones polynomial and the signature. This conjecture was proved by E.-S. Lee [L1]. For arbitrary knots and links, the structure of $\mathcal{H}$ is more complicated than that of the Jones polynomial; right now we do not even have a guess at what the rational homology groups of arbitrary (n, m)-torus knots are.

We next list several interesting applications of $\mathcal{H}$ and related developments.

1) J. Rasmussen used $\mathcal{H}$ and its deformation studied by E. S. Lee [L2] to give a combinatorial proof of the Milnor conjecture that the slice genus of the (p, q)-torus knot is $\frac{(p-1)(q-1)}{2}$ and of its generalization to all positive knots [R2]. This can also be used to show that certain knots are topologically but not smoothly slice without having to invoke Donaldson or Seiberg–Witten gauge theories. Originally, the Milnor conjecture was proved by P. Kronheimer and T. Mrowka via the Donaldson theory [KM].

2) Lenhard Ng [N] obtained an upper bound on the Thurston–Bennequin number of a Legendrian link from its homology $\mathcal{H}(L)$. This bound is sharp on alternating knots and on all but one or two knots with at most 10 crossings.

3) A. Shumakovitch [S] showed that over the 2-element field homology decomposes: $\mathcal{H}(L, \mathbb{F}_2) \cong \widetilde{\mathcal{H}}(L, \mathbb{F}_2) \otimes \mathbb{F}_2[X]/(X^2)$, where $\widetilde{\mathcal{H}}(L, \mathbb{F}_2)$ is the reduced homology of L with coefficients in $\mathbb{F}_2$. P. Ozsváth and Z. Szabó [OS2] discovered a spectral sequence with the E^2-term $\widetilde{\mathcal{H}}(L, \mathbb{F}_2)$ that converges to the Ozsváth–Szabó homology of the double branched cover of L[1].

4) P. Seidel and I. Smith defined a $\mathbb{Z}$-graded homology theory of links via Lagrangian intersection Floer homology of a certain quiver variety [SS]. Their theory is similar to $\mathcal{H}$ in many respects, and, conjecturally, isomorphic to $\mathcal{H}$ after the bigrading in the latter is collapsed to a single grading.

3. Extensions to tangles

The quantum group $U_q(\mathfrak{sl}(2))$ controls the extension of the Jones polynomial to an invariant of tangles, the latter a functor from the category of tangles to the category

of $U_q(\mathfrak{sl}(2))$ representations. To a tangle T with n bottom and m top endpoints (an (m, n)-tangle) there is assigned an intertwiner

$$f(T)\colon V^{\otimes n} \longrightarrow V^{\otimes m}$$

between tensor powers of the fundamental representation V of $U_q(\mathfrak{sl}(2))$.

A categorification of the invariant $f(T)$ was suggested in [BFK]. We considered the category

$$\mathcal{O}^n = \bigoplus_{0\le k\le n} \mathcal{O}^{k,n-k},$$

the direct sum of parabolic subcategories $\mathcal{O}^{k,n-k}$ of a regular block of the highest weight category for $\mathfrak{sl}(n)$. The category $\mathcal{O}^{k,n-k}$ is equivalent to the category of perverse sheaves on the Grassmannian of k-planes in $\mathbb{C}^n$, smooth with respect to the Schubert stratification. The Grothendieck group of $\mathcal{O}^n$ is naturally isomorphic (after tensoring with $\mathbb{C}$) to $V^{\otimes n}$, considered as a representation of $U_{q=1}(\mathfrak{sl}(2))$, and derived Zuckerman functors in $D^b(\mathcal{O}^n)$ lift the action of $\mathfrak{sl}(2)$ on $V^{\otimes n}$. We showed that projective functors in $\mathcal{O}^n$ categorify the action of the Temperley–Lieb algebra on $V^{\otimes n}$ and conjectured how to extend this to arbitrary tangles, by assigning to a tangle T a functor $\mathcal{F}(T)$ between derived categories $D^b(\mathcal{O}^n)$ and $D^b(\mathcal{O}^m)$.

Our conjectures were proved by C. Stroppel [St], who worked with the graded versions $\mathcal{O}^n_{gr}$ of these categories, associated a functor $\mathcal{F}(T)$ between derived categories $D^b(\mathcal{O}^n_{gr})$ and $D^b(\mathcal{O}^m_{gr})$ to each (m, n)-tangle T and a natural transformation $\mathcal{F}(S)\colon \mathcal{F}(T_0) \longrightarrow \mathcal{F}(T_1)$ to a tangle cobordism S between tangles T_0 and T_1. The whole construction is a 2-functor from the 2-category of tangle cobordisms to the 2-category whose objects are $\mathbb{C}$-linear triangulated categories, 1-morphisms are exact functors and 2-morphisms are natural transformations of functors, up to rescalings by invertible complex numbers. When the tangle is a link L, this theory produces bigraded homology groups, conjecturally isomorphic to $\mathcal{H}(L) \otimes \mathbb{C}$.

The braid group action on $V^{\otimes n}$ lifts to a braid group action on the derived category $D^b(\mathcal{O}^n_{gr})$. Restricting to the subcategory $D^b(\mathcal{O}^{1,n-1}_{gr})$ results in a categorification of the Burau representation, previously studied in [KS].

For a more economical extension of the Jones polynomial to tangles, we restrict to even tangles (when the number of endpoints on each of the two boundary planes is even) and to the subspace of $U_q(\mathfrak{sl}(2))$-invariants

$$\mathrm{Inv}(n) = \mathrm{Hom}_{U_q(\mathfrak{sl}(2))}(\mathbb{C}, V^{\otimes 2n})$$

in $V^{\otimes 2n}$. The invariant of a $(2m, 2n)$-tangle is a linear map

$$f_{\mathrm{inv}}(T)\colon \mathrm{Inv}(n) \longrightarrow \mathrm{Inv}(m)$$

between these subspaces.

A categorification of $f_{\mathrm{inv}}(T)$ was found in [K3], [K4]. We defined a graded ring H^n and established an isomorphism

$$K(H^n\text{-mod}) \otimes \mathbb{C} \cong \mathrm{Inv}(n)$$

between the Grothendieck group (tensored with $\mathbb{C}$) of the category of graded finitely-generated H^n-modules and the space of invariants in $V^{\otimes 2n}$. To an even tangle T we assigned an exact functor $\mathcal{T}$ between the derived categories of H^n-mod (this functor induces the map $f_{\mathrm{inv}}(T)$ on the Grothendieck groups) and to a tangle cobordism – a natural transformation of functors. This results in a 2-functor from the 2-category of cobordisms between even tangles to the 2-category of natural transformation between exact functors in triangulated categories. Restricting to links, we recover homology groups $\mathcal{H}(L)$. This approach is more elementary than that via category $\mathcal{O}$, and should carry the same amount of information.

The space of invariants $\mathrm{Inv}(n)$ is a subspace of $V^{\otimes 2n}(0)$, the weight zero subspace of $V^{\otimes 2n}$. A categorification of this inclusion relates rings H^n and parabolic categories $\mathcal{O}^{n,n}$. The latter category is equivalent to the category of finite-dimensional modules over a $\mathbb{C}$-algebra $A_{n,n}$, explicitly described by T. Braden [B]. There exists an idempotent e in $A_{n,n}$ such that $eA_{n,n}e \cong H^n \otimes \mathbb{C}$. This idempotent picks out all self-dual indecomposable projectives in $A_{n,n}$.

Rings H^n can also be used to categorify certain level two representations of $U_q(\mathfrak{sl}(m))$, see [HK].

For a more geometric and refined approach to invariants of tangles and tangle cobordisms we refer the reader to Bar-Natan [BN2]. Some of his generalizations of link homology can be thought of as G-equivariant versions of $\mathcal{H}$, for various compact subgroups G of SU(2), see [K5] for speculations in this direction and for an interpretation of the Rasmussen invariant via the SU(2)-equivariant version of $\mathcal{H}$.

4. $\mathfrak{sl}(n)$ link homology and matrix factorizations

Theorem 4.1. *For each $n > 0$ there exists a homology theory which associates bigraded homology groups*

$$\mathrm{H}_n(L) \cong \bigoplus_{i,j\in\mathbb{Z}} \mathrm{H}_n^{i,j}(L)$$

to every oriented link in $\mathbb{R}^3$. The Euler characteristic of H_n is the polynomial invariant P_n,

$$P_n(L) = \sum_{i,j\in\mathbb{Z}} (-1)^i q^j \dim_{\mathbb{Q}}(\mathrm{H}_n^{i,j}(L)).$$

The homology groups $\mathrm{H}_n^{i,j}(L)$ are finite-dimensional $\mathbb{Q}$-vector spaces, and, for a fixed L, only finitely many of them are non-zero. This homology is functorial: an oriented link cobordism S between L_0 and L_1 induces a homomorphism

$$\mathrm{H}_n(S)\colon \mathrm{H}_n(L_0) \longrightarrow \mathrm{H}_n(L_1),$$

well defined up to overall rescaling by nonzero rationals.

The groups $\mathrm{H}_n(L)$ are constructed in [KR1], where we start with a presentation for $P_n(L)$ as an alternating sum

$$P_n(L) = \sum_{\Gamma} \pm q^{\alpha(\Gamma)} P_n(\Gamma). \tag{1}$$

Here we choose a generic plane projection D of L with m crossings, and form the sum over 2^m planar trivalent graphs Γ which are given by replacing each crossing of D by one of the two planar pictures on the right

Each such planar graph Γ has a well-defined invariant $P_n(\Gamma) \in \mathbb{Z}[q, q^{-1}]$, with all the coefficients being nonnegative integers. Weights $\alpha(\Gamma)$ are given by a simple rule. The edges are of two types: regular oriented edges and "wide" unoriented edges as on the rightmost picture above.

We then define single-graded homology groups $\mathrm{H}_n(\Gamma)$ which have the graded dimension $P_n(\Gamma)$ and satisfy certain naturality conditions allowing us to build a complex out of $\mathrm{H}_n(\Gamma)$, over all modifications Γ of the link diagram D. The complex is a categorification of the right hand side of the equation (1); its homology groups $\mathrm{H}_n(D)$ depend on L only and satisfy the properties listed in Theorem 4.1.

Our definition of $\mathrm{H}_n(\Gamma)$ is based on matrix factorizations. Let $R = \mathbb{Q}[x_1, \ldots, x_k]$. A matrix factorization M of a polynomial $f \in R$ consists of a pair of free R-modules and a pair of R-module maps

$$M^0 \xrightarrow{d} M^1 \xrightarrow{d} M^0$$

such that $d^2 = f \cdot \mathrm{Id}$. The polynomial f is called the *potential* of M. A matrix factorization can be thought of as a two-periodic generalized complex; the square of the differential in not zero, but a fixed multiple of the identity operator. Matrix factorizations were introduced by D. Eisenbud [E] to study homological properties of hypersurface singularities, and later made an appearance in string theory, as boundary conditions in Landau–Ginzburg models [KL]. The tensor product $M \otimes_R N$ of matrix factorizations with potentials f, g is a matrix factorization with potential $f + g$.

To each Γ we associate a collection of matrix factorizations $M_1, \ldots, M_m$, one for each crossing of D, with potentials $f_1, \ldots, f_m$ that add up to zero: $f_1 + f_2 + \cdots + f_m = 0$. The tensor product $M_1 \otimes M_2 \otimes \cdots \otimes M_m$ is a two-periodic complex (since the square of the differential is now zero). Finally, $\mathrm{H}_n(\Gamma)$ is defined as the cohomology of this complex; it inherits a natural $\mathbb{Z}$-grading from that of the polynomial algebra R.

The homology theory H_n is trivial when $n = 1$, while $\mathrm{H}_2(L) \cong \mathcal{H}(L) \otimes \mathbb{Q}$. The theory H_3 should be closely related to the homology theory constructed earlier in [K2]

(the two theories have the same Euler characteristic; the one in [K2] is defined over $\mathbb{Z}$ and not just over $\mathbb{Q}$).

J. Rasmussen [R3] determined homology groups $\mathrm{H}_n(L)$ for all 2-bridge knots L and a few other knots (with a mild technical restriction $n > 4$). Little else is known about homology groups $\mathrm{H}_n(L)$ for $n > 2$.

A Lagrangian intersection Floer homology counterpart of H_n was discovered by C. Manolescu [M]. His theory $\mathcal{H}_{(n)symp}(L)$ is singly-graded, but defined over $\mathbb{Z}$. Manolescu conjectured that $\mathcal{H}_{(n)symp}$, after tensoring with $\mathbb{Q}$, becomes isomorphic to H_n, with the bigrading of the latter folded into a single grading.

5. Triply-graded link homology and beyond

It turns out that the entire HOMFLY-PT polynomial, and not just its one-variable specializations, admits a categorification. The original construction via degenerate matrix factorizations with a parameter [KR2] was later recast in the language of Hochschild homology for bimodules over polynomial algebras [K6]. We represent a link L as the closure of a braid σ with m strands. To σ we assign a certain complex $F(\sigma)$ of graded bimodules over the polynomial algebra R in $m-1$ generators. Taking the Hochschild homology over R of each term in the complex produces a complex of bigraded vector spaces

$$\cdots \longrightarrow \mathrm{HH}(R, F^j(\sigma)) \longrightarrow \mathrm{HH}(R, F^{j+1}(\sigma)) \longrightarrow \cdots$$

The cohomology groups $\mathrm{H}(\sigma)$ of this complex are triply-graded and depend on L only (a convenient grading normalization was pointed out by H. Wu [W]). The Euler characteristic of $\mathrm{H}(L)$ is the HOMFLY-PT polynomial $P(L)$, normalized so that $P(\text{unknot}) = 1$.

This homology theory suffers from two problems. First, the definition requires choosing a braid representative of a knot, rather than just a plane projection. Second, it is not possible to assign maps $\mathrm{H}(S)$ to all link cobordisms S so as to turn H into a functor from LCob to the category of (triply-graded) vector spaces (simply because homology of the unknot is one-dimensional, while that of an unlink is infinite-dimensional). We conjecture that the theory can be redefined on k-component links for all $k > 1$ so as to assign finite-dimensional homology groups $\widetilde{\mathrm{H}}(L)$ to all oriented links L, and not just to knots. The Euler characteristic of $\widetilde{\mathrm{H}}(L)$ will still be the HOMFLY-PT polynomial, but rescaled so as to be a Laurent polynomial in λ and q rather than a rational function. The theory should extend to a projective functor from the category of *connected* link cobordisms to the category of triply-graded vector spaces.

Further extension of $\widetilde{\mathrm{H}}$ to all link cobordisms should only require a minor modification, where one assigns the algebra $\mathbb{Q}[a]$ to the empty link, the differential graded algebra

$$\mathbb{Q}\langle y_1, \ldots, y_n\rangle \otimes \mathbb{Q}[a], \quad y_i y_j + y_j y_i = 0, \quad [y_i, a] = 0, \quad d(y_i) = a, \quad d(a) = 0$$

to a k-component unlink, and suitably resolves each $\widetilde{\mathrm{H}}(L)$, viewed as a $\mathbb{Q}[a]$ module with the trivial action of a, into a complex of free $\mathbb{Q}[a]$-modules.

Understanding $\widetilde{\mathrm{H}}$ could be an important step towards an algebraic description of knot Floer homology, since we expect $\widetilde{\mathrm{H}}$ to degenerate (possibly via a spectral sequence) into knot Floer homology of Ozsváth–Szabó and Rasmussen [OS1], [R1], which categorifies the Alexander polynomial.

An algebraic description of knot and link Floer homology, if someday found and combined with the combinatorial construction [OS3] of Ozsváth–Szabó 3-manifold homology of surgeries on a knot from a filtered version of knot Floer homology (and a generalization of their construction to links), could lead to a combinatorial definition of Ozsváth–Szabó and Seiberg–Witten 3-manifold homology and, eventually, to an algebraic formulation of gauge-theoretical invariants of 4-manifolds.

In conclusion, we mention two other difficult open problems.

I. Categorify polynomial invariants $P(L, \mathfrak{g})$ of knots and links associated to arbitrary complex simple Lie algebras $\mathfrak{g}$ and their irreducible representations.

II. Categorify the Witten–Reshetikhin–Turaev invariants of 3-manifolds.

References

[BN1] Bar-Natan, D., On Khovanov's categorification of the Jones polynomial. *Algebr. Geom. Top.* **2** (2002), 337–370.

[BN2] Bar-Natan, D., Khovanov's homology for tangles and cobordisms. *Geom. Topol.* **9** (2005), 1443–1499.

[BFK] Bernstein, J., Frenkel, I. B., and Khovanov, M., A categorification of the Temperley-Lieb algebra and Schur quotients of $U(\mathfrak{sl}(2))$ via projective and Zuckerman functors. *Selecta Math. (N.S.)* **5** (2) (1999), 199–241.

[B] Braden, T., Perverse sheaves on Grassmannians. *Canad. J. Math.* **54** (3) (2002), 493–532.

[CS] Collin, O., and Steer, B., Instanton Floer homology for knots via 3-orbifolds. *J. Differential Geom.* **51** (1) (1999), 149–202.

[CF] Crane, L., Frenkel, I., Four dimensional topological quantum field theory, Hopf categories, and the canonical bases. *J. Math. Phys.* **35** (1994), 5136–5154.

[E] Eisenbud, D., Homological algebra on a complete intersection, with an application to group representations. *Trans. Amer. Math. Soc.* **260** (1980), 35–64.

[HOMFLY] Freyd, P., Yetter, O., Hoste, Lickorish, W. B. R., K. Millett and A. Ocneanu, A new polynomial invariant of knots and links. *Bull. Amer. Math. Soc. (N.S.)* **12** (2) (1985), 239–246.

[GSV] Gukov, S., Schwarz, A., and Vafa, C., Khovanov-Rozansky homology and topological strings. hep-th/0412243.

[GW] Gukov, S., Walcher, J., Matrix factorizations and Kauffman homology. hep-th/0512298.

[HK] Huerfano, S., and Khovanov, M., Categorification of some level two representations of $\mathfrak{sl}(n)$. math.QA/0204333.

[Ja] Jacobsson, M., An invariant of link cobordisms from Khovanov homology. *Algebr. Geom. Topol.* **4** (2004), 1211–1251.

[J] Jones, V. F. R., A polynomial invariant for knots via von Neumann algebras. *Bull. Amer. Math. Soc.* (*N.S.*) **12** (1985), 103–111.

[KL] Kapustin, A., and Li, Y., D-Branes in Landau-Ginzburg models and algebraic geometry. hep-th/0210296.

[K1] Khovanov, M., A categorification of the Jones polynomial. *Duke Math. J.* **101** (3) (2000), 359–426.

[K2] Khovanov, M., sl(3) link homology. *Algebr. Geom. Topol.* **4** (2004), 1045–1081.

[K3] Khovanov, M., A functor-valued invariant of tangles. *Algebr. Geom. Topol.* **2** (2002), 665–741.

[K4] Khovanov, M., An invariant of tangle cobordisms. *Trans. Amer. Math. Soc.* **358** (2006), 315–327.

[K5] Khovanov, M., Link homology and Frobenius extensions. math.QA/0411447.

[K6] Khovanov, M., Triply-graded link homology and Hochschild homology of Soergel bimodules. math.GT/0510265.

[KR1] Khovanov, M., and Rozansky, L., Matrix factorizations and link homology. math.QA/0401268.

[KR2] Khovanov, M., and Rozansky, L., Matrix factorizations and link homology II. math.QA/0505056.

[KS] Khovanov, M., and Seidel, P., Quivers, Floer cohomology, and braid group actions. *J. Amer. Math. Soc.* **15** (1) (2002), 203–271.

[KM] Kronheimer, P., and Mrowka, T., Gauge theory for embedded surfaces I. *Topology* **32** (4) (1993), 773–826.

[L1] Lee, E. S., The support of the Khovanov's invariants for alternating knots. math.GT/0201105.

[L2] Lee, E. S., An endomorphism of the Khovanov invariant. *Adv. Math.* **197** (2) (2005), 554–586.

[N] Ng, L., A Legendrian Thurston-Bennequin bound from Khovanov homology. *Algebr. Geom. Topol.* **5** (2005), 1637–1653.

[M] Manolescu, C., Link homology theories from symplectic geometry. math.SG/0601629.

[OS1] Ozsváth, P., Szabó, Z., Holomorphic disks and knot invariants. *Adv. Math.* **186** (1) (2004), 58–116.

[OS2] Ozsváth, P., Szabó, Z., On the Heegaard Floer homology of branched double-covers. math.GT/0309170.

[OS3] Ozsváth, P., Szabó, Z., Knot Floer homology and rational surgeries. math.GT/0504404.

[PT] Przytycki, J., and Traczyk,P., Conway algebras and skein equivalence of links. *Proc. Amer. Math. Soc.* **100** (1987), 744–748.

[R1] Rasmussen, J., Floer homology and knot complements. PhD thesis, Harvard University, 2003; math.GT/0306378.

[R2] Rasmussen, J., Khovanov homology and the slice genus. math.GT/0402131.

[R3] Rasmussen, J., Khovanov-Rozansky homology of two-bridge knots and links. math.GT/0508510.

[SS] Seidel, P., and Smith, I., A link invariant from the symplectic geometry of nilpotent slices. *Duke Math. J.*, to appear.

[S] Shumakovitch, A., Torsion of the Khovanov homology. math.GT/0405474.

[St] Stroppel, C., Categorification of the Temperley-Lieb category, tangles, and cobordisms via projective functors. *Duke Math. J.* **126** (3) (2005), 547–596.

[W] Wu, H., Braids, transversal knots and the Khovanov-Rozansky theory. math.GT/0508064.

Department of Mathematics, Columbia University, New York, NY 10027, U.S.A.
E-mail: khovanov@math.columbia.edu

Curve complexes, surfaces and 3-manifolds

Yair N. Minsky*

Abstract. A survey of the role of the complex of curves in recent work on 3-manifolds and mapping class groups.

Mathematics Subject Classification (2000). Primary 57M50; Secondary 30F40, 30F60.

Keywords. Kleinian groups, mapping class group, complex of curves.

1. Disjoint curves in surfaces

The complex of curves, a combinatorial object associated to a surface, has been of interest recently in low-dimensional topology and geometry. In the study of mapping class groups of surfaces, it has shed some light on *relative hyperbolicity* properties, and more generally on the coarse geometry of these groups, and of the Teichmüller spaces (parameter spaces of Riemann surfaces) on which they act. In the setting of hyperbolic 3-manifolds, the complex of curves has been used in the classification theory, particularly the solution of Thurston's Ending Lamination Conjecture, and generally in the attempt to relate more concretely the topology of a 3-manifold to its geometric structure. This paper will attempt to survey some of these developments. Of necessity we will focus on those aspects with which the author is most familiar, thus leaving out a lot of interesting topology and geometry.

We will begin with a leisurely discussion of the natural ways in which simple closed curves (i.e. embedded circles), and particularly the relation of disjointness, occur in low-dimensional topology. After this, in § 2 we will lay out the basic definitions and theorems about curve complexes. In § 3 we will describe work with H. Masur on the inductive structure of curve complexes and its relation to the geometry of mapping class groups. In § 4 we will go into more detail about mapping class groups and outline some recent work with J. Behrstock on their asymptotic cones and the Brock–Farb rank conjecture. In § 5 we will describe work with J. Brock and R. Canary on Thurston's Ending Lamination Conjecture. In § 6 we will lay out some thoughts on the hyperbolic geometry of Heegaard splittings, an area in which our knowledge is still rather incomplete.

*The author is grateful to the NSF for support of this research.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Surfaces and mapping class groups. The first interesting thing one can do with a homotopically essential simple closed curve in a surface is to cut along it, and then glue back. If before gluing back by the identity, the complementary surface is given one full twist in a neighborhood of the curve, the resulting self-map is not homotopic to the identity, and is called a Dehn twist.

The group of orientation preserving homeomorphisms of an oriented surface S to itself, taken modulo isotopy, is called the *mapping class group* of S, or $\mathcal{MCG}(S)$. Dehn twists give rise to infinite cyclic subgroups of $\mathcal{MCG}(S)$, and two disjoint non-homotopic simple curves give rise to commuting Dehn twists.

More generally, let Δ denote a system of disjoint, essential, pairwise non-homotopic simple closed curves (necessarily a finite number, if S is compact). The stabilizer of Δ (up to homotopy) in $\mathcal{MCG}(S)$ is denoted $\mathrm{Stab}(\Delta)$, and we have a short exact sequence

$$0 \to \mathbb{Z}^n \to \mathrm{Stab}(\Delta) \to \mathcal{MCG}'(S \setminus \Delta) \to 0 \tag{1}$$

where $\mathbb{Z}^n$ is the group of Dehn twists generated by elements of Δ and $\mathcal{MCG}'(S \setminus \Delta)$ is the finite-index subgroup of $\mathcal{MCG}(S \setminus \Delta)$ whose elements permute the boundary in such a way that it can still be glued back to obtain Δ in S.

A mapping class that preserves a system of disjoint essential simple closed curves is called *reducible*. Thurston classified the nontrivial conjugacy classes in $\mathcal{MCG}(S)$ as reducible, finite-order, and *pseudo Anosov* [49], [43]. A pseudo-Anosov mapping class does not preserve any finite set of closed curves. Instead, it preserves a pair of *measured geodesic laminations* (see §2), and any closed curve tends, under forward iteration, to one of these and under backward iteration to the other.

The kernel in (1) is an example of an abelian subgroup of $\mathcal{MCG}(S)$. Birman–Lubotzky–McCarthy [13] used Thurston's classification theorem to classify all abelian (and solvable) subgroups of $\mathcal{MCG}(S)$. In particular they showed that the maximal rank of an abelian subgroup is equal to the maximal cardinality of a disjoint system of curves Δ. The pure Dehn-twist group appearing in (1) is not the only way of obtaining a maximal rank abelian group, however – for example if Δ divides S into 3-holed spheres and 1-holed tori, and a Dehn twist is chosen on each component of Δ and a pseudo-Anosov is chosen on each 1-holed torus component, then these elements generate a maximal rank abelian group.

When studying the "coarse" geometry of $\mathcal{MCG}(S)$, i.e. its large-scale geometry when viewed as a metric space by means of its Cayley graph, these maximal abelian subgroups turn out to be quite important. They are quasi-isometrically embedded (see §4 for definitions) by Mosher [105] in the punctured case and Farb–Lubotzky–Minsky [48] in general (See also Theorem 4.1). On the other hand Brock and Farb asked whether these subgroups represent the largest n for which $\mathbb{Z}^n$ is quasi-isometrically embedded in $\mathcal{MCG}(S)$ (not necessarily as a subgroup). This was answered affirmatively by Hamenstädt [58] and by Behrstock–Minsky [8]. The techniques that go into the latter proof are part of what concerns us in this paper.

Hyperbolic geometry. In a complete hyperbolic surface S, each nontrivial homotopy class is represented by a unique geodesic (if it is not homotopic into a cusp). If Δ is a maximal set of disjoint essential simple closed curves, not homotopic to each other or the cusps, then they are realized as a disjoint set of simple geodesics which cut S into 3-holed spheres. It turns out that a hyperbolic 3-holed sphere is determined uniquely by its boundary lengths and all possible lengths can occur. Once lengths are selected for the components of Δ the hyperbolic metric of S is almost determined, except for the gluings of the boundaries back together, for which there is an additional real parameter for each one. This gives *Fenchel–Nielsen coordinates* for the Teichmüller space of S,

$$\mathcal{T}(S) \cong \mathbb{R}_+^{\Delta} \times \mathbb{R}^{\Delta}.$$

The Teichmüller space of S is the set of all hyperbolic metrics on S, up to isometries homotopic to the identity (see e.g. [131], [72], [51]).

On the other hand, the Collar Lemma (see e.g. [78], [38]) asserts that a closed geodesic in S has a regular neighborhood whose radius goes to infinity as the curve's length goes to 0. In particular any sufficiently short geodesic has no self-intersections, and two sufficiently short geodesics cannot cross each other. That suggests a division of $\mathcal{T}(S)$ into *thin regions*, where some curve is shorter than a certain threshold, and *thick regions* where no curves are very short. The intersection pattern of the thin regions is prescribed exactly by the disjointness relation among simple closed curves.

In three dimensional hyperbolic geometry, the generalization of the collar lemma is the "Margulis Lemma" (Kazhdan–Margulis [77]), or Jørgensen's inequality [76] (see also Brooks–Matelski [37], Thurston [131]). In particular sufficiently short closed geodesics have a standard solid torus neighborhood sometimes called a Margulis tube. If for example we consider a hyperbolic structure on $N = S \times \mathbb{R}$, work of Thurston and Bonahon [130], [19] showed that if a geodesic is sufficiently short it must be homotopic to a *simple curve* in S, and Otal [108], [109] showed that any number of sufficiently short curves must be *unknotted and unlinked* – that is, isotopic in N to a collection of disjoint simple level curves with respect to the product structure. See Souto [124] for a generalization of this to the setting of any embedded surface in a hyperbolic 3-manifold.

These sort of results are obtained using *pleated surfaces* (or sometimes *simplicial hyperbolic surfaces*). A pleated surface in a hyperbolic 3-manifold N is a map $f\colon S \to N$ where S is a surface, the pullback path-metric from N gives a complete hyperbolic metric on S, and moreover f is totally geodesic on the complement of a system of simple geodesic lines called a *lamination*. Thurston introduced these tools into hyperbolic geometry, and their great advantage is that (a) their presence greatly constrains the geometry of the 3-manifold in terms of the well-understood geometry of hyperbolic surfaces, and (b) they are plentiful. In fact if $f\colon S \to N$ is a π_1-injective map and Δ is a collection of disjoint, essential, nonhomotopic simple closed curves in S whose f-images are not homotopic to cusps, then there exists a pleated map homotopic to f which carries Δ to its geodesic representatives.

This again gives a geometric interpretation to the disjointness relation for simple curves. Indeed, the Gauss–Bonnet theorem implies that the diameter of a hyperbolic surface is bounded, outside of the thin collars, in terms of its topology. In our 3-manifold setting this means that the geodesic representatives of disjoint curves on S are "close modulo thin parts". Thus the pattern of Margulis tubes in a hyperbolic 3-manifold is closely related to patterns of disjoint curves.

Heegaard splittings. Any closed oriented 3-manifold can be expressed, in infinitely many ways, as a union of two handlebodies glued along their boundaries (a handlebody is a 3-ball with 1-handles attached, or a regular neighborhood of a 1-complex embedded in $\mathbb{R}^3$. For example, consider regular neighborhoods of the 1-skeletons of a triangulation and its dual). This is known as a Heegaard splitting. The non-uniqueness of Heegaard splittings means that it is difficult, although not impossible, to extract meaningful topological information from them. The literature on Heegaard splittings is by now extensive, and our discussion will not come close to being comprehensive. For more information the reader is directed to Birman [14], Scharlemann [116] and Zieschang [133].

Given a surface S, an identification of S with the boundary of a handlebody is determined (up to isotopy) by specifying a maximal set Δ of disjoint nonhomotopic simple curves which are to be the boundaries of essential disks, or *meridians*. (Actually it suffices to choose Δ so that its complement consists of a single genus 0 subsurface). The combination of two such sets for the two handlebodies is called a Heegaard diagram. This is a finite amount of information, but we are faced with the fact that there is no natural choice for this diagram among all possible ones for a given splitting.

A splitting is called *reducible* if there is a curve which is a meridian in both sides. The two disks thus bounded form a sphere, which either bounds a ball, in which case the genus of the splitting can be reduced, or is essential, in which case the manifold can be reduced to a connected sum of manifolds with lower-genus splittings. Haken [56] showed conversely that if a manifold is a connected sum and S is a Heegaard surface, then there exists a sphere that meets S in exactly one meridian. That is, a splitting of a reducible manifold is a reducible splitting. In an irreducible manifold, a splitting of minimal genus is irreducible. However irreducible splittings do not have to be of minimal genus.

Casson–Gordon [44] introduced the notion of *weak reducibility:* A splitting is weakly reducible if it contains two *disjoint* meridians (note that reducible implies weakly reducible, as two equal meridians can be made disjoint). They showed that a weakly reducible splitting is either reducible, or contains a 2-sided incompressible surface (an incompressible surface is a π_1-injectively embedded surface of positive genus; manifolds which admit 2-sided incompressible surfaces are called *Haken*, and are in many ways easier to study.) Thus an irreducible non-Haken manifold has a *strongly irreducible* (i.e. not weakly reducible) splitting. For Haken manifolds, extensions of the arguments of Casson–Gordon by Scharlemann–Thompson (see [117] and [115])

lead to a decomposition along incompressible surfaces into strongly irreducible Heegaard splittings-with-boundary, known as a *generalized Heegaard splitting*. This has become a widely applied and sophisticated theory.

Hempel [69] showed that a splitting of a manifold that is Seifert-fibred or contains an essential torus must have the *disjoint curve property*: There exist meridians m_1, m_2 from the two sides, and a curve γ so that both m_1 and m_2 are disjoint from γ. (Note that reducibility and weak-reducibility are special cases of this property).

At this point, the properties of splittings and their meridians come into contact with Thurston's geometrization conjecture. After decomposing along essential spheres and tori, a 3-manifold should fall into pieces admitting one of the eight 3-dimensional geometries – Euclidean, hyperbolic, spherical, or the five non-isotropic fibred geometries (see Scott [122]). Our discussion so far implies that manifolds that break up non-trivially have reducible splittings (by Haken) or splittings with the disjoint curve property (by Hempel).

Irreducible splittings for the non-hyperbolic pieces are now very well understood, through the work of Waldhausen [132], Bonahon–Otal [20], Moriah [103], Boileau–Rost–Zieschang and Boileau–Collins–Zieschang [17], [15], Frohman–Hass [50], Boileau–Otal [16], Moriah–Schultens [104] and Cooper–Scharlemann [46]. In all these cases splittings have the disjoint curve property.

A splitting without the disjoint curve property should therefore yield a hyperbolic manifold (modulo the geometrization conjecture, whose proof by Perelman is, as of this writing, edging ever closer to formal acceptance). Hempel brought the notions of reducibility, weak reducibility, and the disjoint curve property together as particular values of a *distance function on splittings*, which we will discuss further in §6. The main question that we will discuss in §6 is: how do we translate information in the Heegaard diagram into geometric data for a hyperbolic manifold?

2. Curve complexes

In view of the foregoing discussion, one is naturally led to consider a combinatorial object which captures the disjointness relation among simple closed curves (up to homotopy) on a surface. Actually when Harvey introduced the complex of curves in [63], [64] he was motivated by analogy, in the setting of $\mathcal{MCG}(S)$, with Bruhat–Tits buildings for Lie groups. We will now give some precise definitions, which are somewhat more tedious than one would hope because of the need to deal with a few special cases.

2.1. Definitions. Let $S = S_{g,b}$ be an orientable compact surface with genus g and b boundary components. An *essential simple curve* in S will, for us, be an embedded circle in S which is homotopically non-trivial, and not homotopic into ∂S (non-peripheral).

Let $\mathcal{C}(S)$ denote the simplicial complex whose vertices are homotopy classes of essential simple curves on S, and whose k-simplices are, except in a few sporadic cases described below, defined to be $(k+1)$-tuples of distinct vertices $[\alpha_0, \ldots, \alpha_k]$ whose representatives can be chosen to be pairwise disjoint. One may check that, when S has negative Euler characteristic, that there are at most $\xi(S) \equiv 3g - 3 + b$ such curves, and so $\dim \mathcal{C}(S) = \xi(S) - 1$. Although $\mathcal{C}(S)$ is finite dimensional, it is not locally finite, and this complexity accounts for much of the interest in studying and applying it. We let $\mathcal{C}_k(S)$ denote the k-skeleton of $\mathcal{C}(S)$.

Sporadic cases. For $S_{0,b}$ with $b \leq 3$ the complex is empty (although the annulus $S_{0,2}$ will later play an important role in a different way). For the tori $S_{1,0}$ and $S_{1,1}$, and for the 4-holed sphere $S_{0,4}$, the above definition gives a 0-dimensional complex. It turns out to be useful to add edges according to the rule that $[vw]$ is an edge if v and w have representatives that intersect once (in the case of the tori) or twice (in the case of the sphere). This forms our definition of $\mathcal{C}(S)$ for these cases.

Examples. For $S_{1,0}$ and $S_{1,1}$, simple curves are identified by their slopes in homology, i.e. $\mathcal{C}_0(S) \cong \mathbb{Q} \cup \{\infty\}$. Edges correspond to edges of the *Farey triangulation* of the disk, that is $\left[\frac{p}{q}, \frac{r}{s}\right]$ is an edge iff $|ps - qr| = 1$ (with fractions in lowest terms). See Figure 1. $\mathcal{C}(S_{0,4})$ is also isomorphic to $S_{1,1}$, with the isomorphism obtained from the fact that $S_{1,0}$ or $S_{1,1}$ modulo the hyperelliptic involution, and punctured at the fixed points, is $S_{0,4}$.

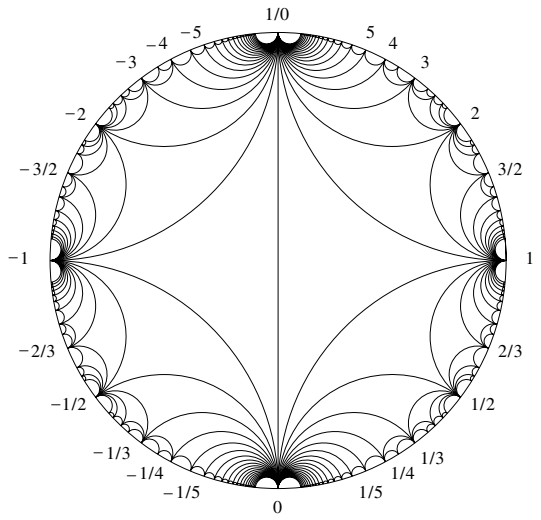


Figure 1. The Farey triangulation is the curve complex of $S_{1,0}$, $S_{1,1}$ and $S_{0,4}$.

For $S = S_{0,5}$ and $S_{1,2}$ (the case $\xi(S) = 2$), $\mathcal{C}(S)$ is again a graph, and in fact the two are isomorphic (again via the hyperelliptic involution). The link of a vertex $v \in \mathcal{C}(S_{0,5})$ is $\mathcal{C}_0(S_{0,4}) = \mathbb{Q} \cup \{\infty\}$.

Mapping class group action. $\mathcal{MCG}(S)$ acts naturally on $\mathcal{C}(S)$ and since homeomorphisms preserve disjointness it acts by simplicial automorphisms, and the quotient is easily seen to be finite (the set of vertices $[\alpha]$ of $\mathcal{C}(S)/\mathcal{MCG}(S)$, for example, is indexed by the topological type of $S \setminus \alpha$). The action is far from proper, as the large stabilizer in (1) indicates (we now recognize Δ as a simplex in $\mathcal{C}(S)$).

Nevertheless the simplicial structure of $\mathcal{C}(S)$ records all of the structure of $\mathcal{MCG}(S)$ in the following sense:

Theorem 2.1 (Ivanov [75], Luo [86], Korkmaz [82]). *The map $\mathcal{MCG}(S) \to \mathrm{Aut}(S)$ is an isomorphism in all cases except for $S_{1,2}$, where it is injective with index 2 image.*

(Ivanov proved this for genus at least 2. Luo and Korkmaz proved the remaining cases, with Luo's proof giving a unified argument.)

Hatcher–Thurston [67] defined a closely related complex and used it to give a presentation for $\mathcal{MCG}(S)$. Harer [61], [60] computed the homotopy type of $\mathcal{C}(S)$ and related complexes, and used it to study the homology of $\mathcal{MCG}(S)$. Hatcher [65] gave simplified proofs of contractibility for related complexes of arcs in a punctured surface.

2.2. Geometric structure. As the local structure of $\mathcal{C}(S)$ is rather intricate, it turns out that something can be gained by attempting to ignore it. First let us view $\mathcal{C}(S)$ as a metric space by making each simplex standard Euclidean with sidelength 1, say, and taking the path metric (a theorem of Bridson [24] assures us that this makes it into a complete geodesic metric space). It will sometimes be simpler to consider the path metric just on the 1-skeleton $\mathcal{C}_1(S)$. These two metric spaces are quasi-isometric, and we will be interested in coarse features such as quasi-geodesics, quasi-isometric type, etc.

Hyperbolicity. With Masur in [94], we proved the following theorem:

Theorem 2.2. *In all nontrivial cases $\mathcal{C}(S)$ is an infinite diameter δ-hyperbolic metric space.*

The usual definition of δ-hyperbolic, due to Rips, Gromov and Cannon, is equivalent for complete path metric spaces to the *δ-thin triangles property:* each leg of a geodesic triangle is contained in a δ-neighborhood of the other two legs.

Examples. Typical examples to keep in mind are metric trees, classical hyperbolic spaces, and Cayley graphs of fundamental groups of closed negatively curved manifolds. $\mathbb{R}^n$ and $\mathbb{Z}^n$ are not δ-hyperbolic when $n > 1$, and in fact any group that contains a $\mathbb{Z}^2$ subgroup does not have a δ-hyperbolic Cayley graph. Note, this includes $\mathcal{MCG}(S)$ for non-sporadic S.

The proof of Theorem 2.2 involved the construction of a family of paths in $\mathcal{C}(S)$ using geodesics in the Teichmüller space of S. If to each point along a Teichmüller

geodesic we associate the set of shortest curves in the associated metric, we obtain a "quasi-path" in $\mathcal{C}(S)$. Properties of quadratic differentials and their foliations were used to construct a "quasi-retraction" from $\mathcal{C}(S)$ to such a path, which has certain strong contraction properties that imply that $\mathcal{C}(S)$ is hyperbolic, and the paths are quasi-geodesics.

See Bowditch [22] for a considerable simplification of this proof, which in particular provides explicit upper bounds for δ, logarithmic in $\xi(S)$.

Laminations and boundary. A δ-hyperbolic space has a natural boundary at infinity, whose points are equivalence classes of "fellow-traveling" quasigeodesic rays. Klarreich's theorem describes this boundary, and relates it to Thurston's theory of geodesic laminations on surfaces.

Fix a complete, finite-area hyperbolic metric on $\mathrm{int}(S)$. A geodesic lamination on a complete hyperbolic surface is a closed set which is a disjoint union of complete simple geodesics. For example, a closed geodesic loop is a lamination, and a sequence of such loops whose lengths increase without bound will have a subsequence that converges on compact sets to a geodesic lamination, in the Hausdorff topology. Thurston used this notion to complete the set of geodesic loops in S. (His construction involved additional structure – he considers laminations equipped with *transverse measures* – but we will ignore this point here).

We note also that different choices of hyperbolic structures give rise to canonically homeomorphic representations of the space of geodesic laminations, so we may consider this as an intrinsic topological object (see Hatcher [66]).

Of interest to us will be the set of *filling* laminations, which we denote $\mathcal{EL}(S)$. A lamination is filling if it intersects every simple closed geodesic. $\mathcal{EL}(S)$ comes with a natural topology (which is somewhat coarser than the topology of Hausdorff convergence, and involves the transverse measures).

Example. For $S_{1,1}$ and $S_{0,4}$ we have seen that $\mathcal{C}_0(S)$ can be identified with $\mathbb{Q} \cup \{\infty\}$. The set of all laminations is the circle $\mathbb{R} \cup \infty$, with irrational slopes corresponding to filling laminations.

Klarreich's theorem states the following:

Theorem 2.3 (Klarreich [79]). *The boundary of $\mathcal{C}(S)$ is naturally homeormorphic to $\mathcal{EL}(S)$.*

The proof uses the same Teichmüller-geodesic paths used in the proof of hyperbolicity. A Teichmüller ray has a "foliation at infinity" whose length shrinks exponentially as one travels along the ray. These foliations correspond to the laminations in $\mathcal{EL}(S)$ (see Levitt [85] for the general correspondence between foliations and laminations). An alternate proof was found by Hamenstädt [59], using the machinery of train-tracks.

The topological structure of $\mathcal{EL}(S)$ is still somewhat mysterious. It is not known for example if it is disconnected, except in the case $\xi = 1$ where it is the set of irrationals in $\mathbb{R}$ (this question was first raised by P. Storm).

3. Nested structure

The coarse information given by the hyperbolicity theorem is hard to use by itself. With Masur in [95], we refine the coarse approach by applying it to the link structure of $\mathcal{C}(S)$. The link of a simplex in $\mathcal{C}(S)$ is closely related to the curve complex of a subsurface of S, and hence the idea here is to use coarse geometry in an inductive way, looking successively at finer levels of structure.

Subsurfaces and partitions. Let W be an essential subsurface of S, and suppose first that $\xi(W) > 1$. It is evident that $\mathcal{C}(W)$ is a subcomplex of $\mathcal{C}(S)$, contained in the link of the simplex $[\partial W]$. The remaining cases, namely when W is an annulus, one-holed torus and four-holed sphere, cause a certain amount of extra trouble. For $W = S_{1,1}$ and $W = S_{0,4}$, for example, all the vertices of $\mathcal{C}(W)$ are vertices of $\mathcal{C}(S)$, but the edges are not edges of $\mathcal{C}(S)$, due to the special definition of $\mathcal{C}(W)$ in that case. When W is an annulus, $\mathcal{C}(W)$ is empty but we will want something that will correspond to the Dehn twist group around W. (If $W = S_{0,3}$ then $\mathcal{C}(W) = \emptyset$ also, and we will be happy with that).

So, let us make the following new definitions. Let W be a surface with nonempty boundary. Let $\mathcal{A}(W)$ be the complex whose vertices are essential curves (i.e. vertices of $\mathcal{C}(W)$) and properly embedded essential arcs. Except when W is an annulus, we consider these things up to isotopy rel boundary, and "essential" for arcs means not isotopic into the boundary. For an annulus we take the isotopies to have fixed endpoints. Simplices are sets of arcs and curves that have representatives with disjoint interiors. This makes $\mathcal{A}$ infinite dimensional for an annulus, but it is still quasi-isometric to $\mathbb{Z}$. For $\xi(W) > 1$, we note that $\mathcal{C}(W)$ embeds in $\mathcal{A}(W)$ and this embedding is a quasi-isometry (when $\xi(W) = 1$ there is a quasi-isometry which is not quite an embedding).

If Δ is a simplex of $\mathcal{C}(S)$, let $\sigma(\Delta)$ be the union of components of $S \setminus \Delta$ which are not 3-holed spheres, together with annuli whose cores are components of Δ. This is called the "partition" of Δ. We let $lk^*(\Delta)$ denote the join of $\mathcal{A}(W_i)$ over all components $W_i \in \sigma(\Delta)$.

We think of $lk^*(\Delta)$ as an "extended link" for Δ. The actual link of Δ is the join of the complexes for components of $S \setminus \Delta$ of $\xi > 1$, and the 0-skeletons of the complexes for $\xi = 1$ components. The annular factors are not visible in the link; they can be detected in the neighborhood of radius 2, however.

Basic geometric properties. There is a small number of geometric properties of $\mathcal{C}(S)$ which connect global geometry to geometry in the extended links lk^*, and are responsible for most of the rest of our analysis. In outline they are the following:

1. *Hyperbolicity at all levels.* By Theorem 2.2 $\mathcal{C}(S)$ is hyperbolic, and furthermore for every simplex Δ, $lk^*(\Delta)$ is the join of the hyperbolic complexes associated to the components of $\sigma(\Delta)$.

2. *Subsurface projection bounds.* There is a natural projection from $\mathcal{C}(S)$ minus a neighborhood of $[\partial W]$ to $\mathcal{A}(W)$, where W is an essential subsurface. A geodesic in $\mathcal{C}_1(S)$ which stays out of this neighborhood has uniformly bounded projection image.
3. *Hierarchy paths and rigidity.* There is a distinguished family of quasigeodesic paths in the graph of markings on S (see below) which are controlled in a strong way by the subsurface projections of their endpoints.

Subsurface projections. For any essential $W \subset S$ there is a useful map

$$\pi_W : \mathcal{C}(S) \to \mathcal{A}(W) \cup \{\emptyset\}.$$

Namely, given a point $x \in \mathcal{C}(S)$, let δ be the simplex whose interior contains x. There is a unique cover of S to which W lifts homeomorphically, and this cover has a natural compactification $\overline{W}$ that identifies it with W (inherited from the natural compactification of the universal cover of S). Each component of δ lifts to a system of curves and/or arcs in $\overline{W}$, and the essential ones define vertices of $\mathcal{A}(W)$. The union over all vertices of δ gives a (possibly empty) simplex in $\mathcal{A}(W)$, and barycentric coordinates of x in δ push forward to define a unique point in this simplex, which is $\pi_W(x)$. Note that $\pi_W(x) = \emptyset$ only if δ has no essential intersections with W. (See also Ivanov [74] for a version of π_W).

One can think of this map as analogous to visual projection from $X \setminus \{x\}$ to the unit sphere around x, for a reasonably nice space X.

Link projection bounds. In [95], we proved the following theorem:

Theorem 3.1. *Let $W \subset S$ be an essential surface. let g be a geodesic in the 1-skeleton of $\mathcal{C}(S)$ all of whose vertices intersect W essentially. Then*

$$\operatorname{diam}_{\mathcal{A}(W)}(g) \leq B$$

where B depends only on the topological type of S.

A motivational analogy for this theorem may be found in the setting of CAT(0) complexes. (CAT(0) refers to non-positive curvature in the sense of comparison geometry, and in particular implies uniqueness of geodesics between points. See e.g. [25].) Let v be a point in a piecewise Euclidean CAT(0) complex X, and let $l(v)$ be its link, which we may identify with the unit "sphere" around v. There is a well-defined projection of $X \setminus \{v\}$ to $l(v)$, via geodesic segments from points in X to v, and the diameter $d_v(Y)$ of the projection of $Y \subset X \setminus \{v\}$ is the "visual diameter of Y". If g is a geodesic segment that avoids v then the cone of g onto v is an embedded triangle, and it follows that $d_{l(v)}(g) < \pi$. Conversely if $d_{l(v)}(g) \geq \pi$ it follows that g must pass through v.

The reason that this theorem holds in the curve complex, in spite of there not being a CAT(0) metric, is roughly the following: Let g be the geodesic and imagine that g

is extended to an infinite ray $\widehat{g}$, still with all vertices crossing W essentially (this is always possible in one direction or the other). Klarreich's theorem implies that $\widehat{g}$ converges to a filling lamination λ. The lift of λ to $\widehat{W}$ gives a point (or really simplex) in $\mathcal{A}(W)$, and since the vertices of $\widehat{g}$ converge to λ in a geometric sense, eventually their projections to $\mathcal{A}(W)$ are within bounded distance of the projection of λ. Thus $\operatorname{diam}_{\mathcal{A}(W)}(\widehat{g}) < \infty$.

Now to obtain the bound of Theorem 3.1, we need a uniform version of the above argument. This requires a more delicate analysis using some of the same machinery used to prove the hyperbolicity theorem.

Theorem 3.1 is the first step in a finer inductive study of the geometry of $\mathcal{MCG}(S)$. In order to describe this, we need to introduce some more terminology.

Markings and hierarchies. A convenient way to study the geometry of $\mathcal{MCG}(S)$ is to look at its action on the *marking graph* $\mathcal{M}(S)$. A marking μ of S is given by the following data: A maximal simplex of $\mathcal{C}(S)$, called base(μ), and a collection of *transversal curves* trans(μ), where each base curve α is equipped with one transversal curve t_α, which intersects α. t_α is disjoint from all other base curves, and α and t_α either intersect exactly once, or twice with opposite orientations (their regular neighborhood is then, respectively, $S_{1,1}$ or $S_{0,4}$). It is easy to see that $\mathcal{MCG}(S)$ acts on these markings with finite quotient and finite stabilizers. Moreover one can easily write down a simple finite list of "elementary moves" $\mu \to \nu$ such that the graph $\mathcal{M}(S)$ whose vertices are markings and whose edges are elementary moves is connected. Hence $\mathcal{M}(S)$, with the natural metric in which every edge has unit length, is quasi-isometric to $\mathcal{MCG}(S)$, which acts on it properly, cocompactly and isometrically.

In [95], we study a class of paths in $\mathcal{M}(S)$ which arise from an iterated construction in $\mathcal{C}(S)$ which we call a "hierarchy of geodesics".

Let us begin with an example (the same one treated extensively in [101]). If $S = S_{0,5}$ then $\mathcal{C}(S)$ is a graph, and a marking consists of two base curves and two transversals. Let μ and ν be two markings. Let $v_0, v_1, \ldots, v_N$ be the vertices of a geodesic in $\mathcal{C}_1(S)$ joining $v_0 \in \text{base}(\mu)$ to $v_N \in \text{base}(\nu)$. For $0 < i < N$ we note that v_i cuts S into $W_i \cong S_{0,4}$ and a three-holed sphere. Both v_{i-1} and v_{i+1} are vertices in $\mathcal{C}(W_i)$, and we may join them by a geodesic $v_{i-1} = u_0, \ldots, u_k = v_{i+1}$ in $\mathcal{C}_1(W_i)$. This gives us a sequence

$$[v_i, u_0], \ldots, [v_i, u_k]$$

of edges, or pants decompositions of S, with each step corresponding to a simple curve-replacement move. At the beginning we similarly join the second vertex of base(μ) to v_1 in W_0, and likewise at the end with base(ν). The result of this is a sequence of pants decompositions of S joining μ to ν, and separated by simple moves. Now consider an interior point u_j of the $\mathcal{C}(W_i)$-geodesic built over v_i. u_{j-1} and u_{j+1} intersect u_j, and they give two points in $\mathcal{A}(u_j)$, the annulus complex. We can join one to the other by a geodesic in $\mathcal{A}(u_j)$, which amounts to a sequence of Dehn twists. Something slightly subtle happens at the endpoints of the $\mathcal{C}(W_i)$-geodesics, which the

reader is invited to investigate. This procedure extends all the pants decompositions we built into markings, which can be traversed in a sequence of elementary moves from μ to ν.

In general something similar happens – starting with a geodesic in $\mathcal{C}_1(S)$ we inductively add geodesics in the extended links of the simplices traversed. (Actually we work with sequences of simplices called "tight geodesics", but we will ignore this technicality here). The final output of this procedure is not exactly a sequence of markings; this is because, when the process fills in two or more disjoint subsurfaces, the geodesics in those subsurfaces can be traversed in either order. That is, we are dealing with the unavoidable appearance of product regions in $\mathcal{M}(S)$. However the structure can be resolved (non-uniquely) into a sequence of markings connected by elementary moves. We call these *hierarchy paths*. The following theorem summarizes the properties of hierarchies and their paths established in [95]:

Theorem 3.2. *For any pair of points $\mu, \nu \in \mathcal{M}(S)$ there is a hierarchy and a family of hierarchy paths with the following properties.*

- Efficiency. *Hierarchy paths are quasigeodesics in $\mathcal{M}(S)$, with uniform constants.*
- Monotonicity. *For any essential $W \subseteq S$ and a hierarchy path β, $\pi_W \circ \beta$ traverses quasi-monotonically a bounded neighborhood of a geodesic between $\pi_W(\mu)$ and $\pi_W(\nu)$.*
- Forced traversals. *There is a constant B depending only on the topology of S such that if, for $W \subset S$, we have $d_W(\mu, \nu) > B$, then W appears in any hierarchy for μ and ν, and in particular any hierarchy path must have a subpath in which all the markings contain $[\partial W]$.*
- Partial ordering and stability. *There is a partial order defined among subsurfaces which appear in a hierarchy, such that any two subsurfaces which intersect essentially are ordered. If $W < Z$ in this order for a hierarchy from μ to ν then ∂W appears before ∂Z in any hierarchy path from μ to ν. If also $d_W(\mu, \nu)$ and $d_Z(\mu, \nu)$ are larger than a certain a priori constant, then this partial ordering is consistent over all hierarchies from μ to ν, and moreover for any ν' for which Z appears in a hierarchy from μ to ν', W is forced to appear as well.*

(We have abbreviated $d_{\mathcal{A}(W)}(\pi_W(\mu), \pi_W(\nu))$ as $d_W(\mu, \nu)$, and will continue to do this.) The Forced Traversals property is essentially a generalization of Theorem 3.1.

Distance formula. The partial ordering and monotonicity properties mean that the length of a hierarchy path is, roughly, the sum of the projection distances of its endpoints in all subsurfaces which it traverses. Moreover the forced traversals property implies that those subsurfaces traversed account for all subsurfaces in which these projection distances are sufficiently large. Any competing path from μ to ν is forced

to make up the same distances at some point, and this is the basic reason for the quasi-geodesic property of hierarchy paths. This argument also gives rise to the following Distance Formula, which plays an important role later on.

Theorem 3.3. *If $\mu, \nu \in \mathcal{M}(S)$,*

$$d_{\mathcal{M}(S)}(\mu, \nu) \approx \sum_{Y \subseteq S} \{d_Y(\mu, \nu)\}_K . \tag{2}$$

Some explanations are in order here: We define the expression $\{N\}_K$ to be N if $N > K$ and 0 otherwise – hence K functions as a "threshold" below which contributions are ignored. The constant K used depends only on the topological type of S, and the expression $f \approx g$ means $g/a - b < f < ag + b$ where $a > 1, b > 0$ are also a priori constants. Note that, if $W \subset S$ is an essential surface and $\mu \in \mathcal{M}(S)$ is a marking, then $\pi_W(\mu) \in \mathcal{A}(W)$ is always defined (up to finitely many choices), since some part of μ, possibly only a transversal curve, must intersect W essentially.

So this theorem is saying that, after throwing away the low-level "noise", only finitely many subsurfaces $W \subseteq S$ remain, and the projection distances in these account for the distance between μ and ν. Note also that the sum includes a term for $Y = S$, i.e. distance in the curve complex itself.

Geodesics in Teichmüller space. Just as Teichmüller geodesics play a role in the proof of hyperbolicity for $\mathcal{C}(S)$, the geometry of $\mathcal{C}(S)$ and its nested structure gives us some added understanding of Teichmüller geodesics. Rafi [112], [111] analyzed the extent to which the long-term behavior of a Teichmüller geodesic mirrors the combinatorial structure of a hierarchy, and also develops in [110] a distance formula in $\mathcal{T}(S)$ analogous to Theorem 3.3. See also Rees [113] for a related but independent study of Teichmüller geodesics.

4. Coarse geometry of $\mathcal{MCG}(S)$

The study of coarse geometric properties of abstract groups, by means of their Cayley graphs (or equivalently their word metric), can be traced back to the theorem of Milnor and Švarc [98], [128] on growth rates of groups, to Gromov's work [53] on groups of polynomial growth, and the introduction by Gromov [54] and Cannon [41], [42] of hyperbolic groups. This field is now enormous and we will not attempt to survey it. Some good general references are Gromov [55], [54] and Bridson–Haefliger [25].

We will be interested in examining phenomena of hyperbolicity, undistorted (quasi-isometrically embedded) subgroups, geometric rank, and asymptotic cones, as they relate to $\mathcal{MCG}(S)$. $\mathcal{MCG}(S)$ is not hyperbolic, but as we have already seen it is hyperbolic in a relative sense, through its action on the hyperbolic space $\mathcal{C}(S)$. Before we proceed let us record some of the usual definitions.

A map $f\colon X \to Y$ between metric spaces is *coarse Lipschitz* if a uniform inequality holds of the form

$$d(f(x), f(x')) \leq ad(x, x') + b$$

for all $x, x' \in X$. It is *coarsely bilipschitz*, or a *quasi-isometric embedding*, if the opposite inequality

$$d(x, x') \leq ad(f(x), f(x')) + b$$

holds as well. Finally we say that $f\colon X \to Y$ is a *quasi-isometry* if it is coarsely bilipschitz, and in addition there is an upper bound on $d(y, f(X))$ for all $y \in Y$.

We assume from now on that generators for $\mathcal{MCG}(S)$ are fixed, giving $\mathcal{MCG}(S)$ a word metric, which (as remarked in § 3) is quasi-isometric to $\mathcal{M}(S)$.

Quasi-isometrically embedded subgroups. One almost immediate consequence of the Distance Formula is this theorem:

Theorem 4.1. *Let Δ be a simplex in $\mathcal{C}(S)$. The subgroup* Stab(Δ) *is quasi-isometrically embedded in $\mathcal{MCG}(S)$. Moreover, there is a coarse-Lipschitz retraction $\mathcal{MCG}(S) \to$* Stab(Δ). *Finally,* Stab(Δ) *is quasi-isometric to a product*

$$\prod_{W \in \sigma(\Delta)} \mathcal{MCG}(W)$$

where $\sigma(\Delta)$ refers to the partition of S defined by Δ, as in § 3.

(For annular components of $\sigma(\Delta)$, we interpret $\mathcal{MCG}(W)$ to be the Dehn twist group of W).

The constants of the quasi-isometry and retraction depend on the choice of Δ. A uniform statement is obtained by fixing Δ and considering all left-cosets of Stab(Δ). More geometrically, and more in keeping with our approach, we consider the marking graph $\mathcal{M}(S)$, and within it the subsets $Q(\Delta)$ which consist of all markings whose base contains the simplex Δ. This subset is clearly quasi-isometric to Stab(Δ) and its left cosets. With this notation, what we produce is a coarse-Lipschitz retraction $\mathcal{M}(S) \to Q(\Delta)$, and a quasi-isometry

$$Q(\Delta) \cong \prod_{W \in \sigma(\Delta)} \mathcal{M}(W). \tag{3}$$

Where now the constants depend only on the topological type of S, and not on the choice of Δ. For an annulus W, $\mathcal{M}(W)$ is just $\mathcal{A}(W)$, which we recall is quasiisometric to $\mathbb{Z}$.

Example. If $S = S_{0,5}$ and $\Delta = [\delta_0, \delta_1]$ is an edge, then $Q(\Delta)$ is quasi-isometric to $\mathbb{Z}^2$ (and stabilized by the $\mathbb{Z}^2$ subgroup of Dehn twists about δ_0 and δ_1). If Δ is a vertex then $Q(\Delta)$ is quasi-isometric to $\mathbb{Z} \times \mathrm{SL}_2(\mathbb{Z})$, with the second factor corresponding to $\mathcal{M}(S_{0,4})$. Compare to the short exact sequence (1).

The quasi-isometric embedding part of the statement is a reflection of the fact that in the distance formula for two elements of $Q(\Delta)$, the only contributions come from terms that have no essential intersection with Δ, i.e. those that contribute also to the product over $\sigma(\Delta)$. See also Hamenstädt [57] for a proof of a similar statement using train-track technology.

The coarse retraction to $Q(\Delta)$ is built, inductively, using the subsurface projections described in the previous section. See Behrstock [7] for details. We call this retraction $\pi_{Q(\Delta)}$. The composition of $\pi_{Q(\Delta)}$ with projection to any of the factors $\mathcal{M}(W)$ of the product in (3) is called $\pi_{\mathcal{M}(W)}$, and it is well defined up to bounded ambiguity.

Asymptotic cones. Another way to quantify our coarse understanding of the group is to consider its *asymptotic cones*, which are rescaling limits of the group, in a certain sense. They were used implicitly by Gromov in [53], and introduced explicitly in Van den Dries–Wilkie [47]. In order to define the limit one resorts to the mechanism of *ultrafilters*, which can be briefly described as follows.

Let $X = (X, d)$ be a metric space, fix a sequence $s_n \to \infty$, and consider the metric spaces $(X_n, d_n) \equiv (X, \frac{1}{s_n} d)$, namely $X$with metric scaled down by $1/s_n$. We want, essentially, to consider the set of all "limiting pictures" of X_n, as $n \to \infty$. An *ultrafilter* is a way of organizing the natural numbers so as to pick out a convergent subsequence for every sequence of points in a compact space. More precisely, let ω be a *finitely additive probability measure* on $\mathbb{N}$, which is defined on every subset, and takes on only values of 0 and 1. Further, we assume ω is *non-principal*, meaning it is 0 on finite subsets. Existence of such measures is a nice exercise in using Zorn's lemma. If p_n is a sequence in a Hausdorff space T, we declare the ω-limit to be $\lim_\omega p_n = p$ if, for every neighborhood U of p, $\omega\{j : x_j \in U\} = 1$.

With this definition we find that *every* sequence in a compact space has a unique ω-limit. Now returning to X_n, we consider all sequences $\boldsymbol{x} = (x_n \in X_n)$. A pseudo-distance can be defined by

$$d_\omega(\boldsymbol{x}, \boldsymbol{y}) = \lim_\omega d_n(x_n, y_n)$$

which gives a point in $[0, \infty]$. Fixing a basepoint sequence $\boldsymbol{x}_0$ we can restrict to the subset $\{\boldsymbol{x} : d_\omega(\boldsymbol{x}_0, \boldsymbol{x}) < \infty\}$ and identify pairs $\boldsymbol{x} \sim \boldsymbol{y}$ whenever $d_\omega(\boldsymbol{x}, \boldsymbol{y}) = 0$. This gives a metric space, known as an asymptotic cone for X.

Note that this construction depends on the choice of scaling constants, ultrafilter ω, and basepoints. In our setting X will be $\mathcal{M}(S)$ (or equivalently $\mathcal{MCG}(S)$), and the basepoint will not matter because the space is (coarsely) homogeneous. The scaling constants and ultrafilter will be assumed fixed. We denote the asymptotic cone of $\mathcal{M}(S)$ by $\mathcal{M}^\omega(S)$.

Examples. The asymptotic cone of $\mathbb{Z}^n$ is always $\mathbb{R}^n$. The symptotic cone of a δ-hyperbolic space is an $\mathbb{R}$-tree, that is, a geodesic metric space in which any two points are joined by a unique embedded path. $\mathcal{MCG}(S_{1,1})$ and $\mathcal{MCG}(S_{0,4})$ are both commensurable with $\mathrm{SL}_2(\mathbb{Z})$, and hence with the free group F_2 whose Cayley graph

is a tree. Hence their asymptotic cones are $\mathbb{R}$-trees as well. (Note that these $\mathbb{R}$-trees have dense, and uncountable, branching).

A quasi-isometric embedding gives rise, after taking asymptotic cones, to a bilipschitz embedding. This allows us to replace coarse statements with topological statements, which is part of what makes asymptotic cones useful.

A sequence of subsets $Q(\Delta_i)$ of $\mathcal{M}(S)$ gives rise to an ω-limit which we denote by $\mathcal{Q}^\omega(\mathbf{\Delta}) \subset \mathcal{M}^\omega(S)$ (where $\mathbf{\Delta}$ denotes the sequence (Δ_i)). Theorem 4.1, or rather the quasi-isometry (3), immediately tells us that $\mathcal{Q}^\omega(\mathbf{\Delta})$ is a bilipschitz-embedded product of lower-complexity asymptotic cones, admitting a Lipschitz retraction from $\mathcal{M}^\omega(S)$.

This retraction and some related constructions are instrumental in studying separation and dimension properties of the asymptotic cone, as is done in Behrstock [7] and Behrstock–Minsky [8]. The main theorem of [8] is the following:

Theorem 4.2. $\widehat{\dim}\, \mathcal{M}^\omega(S) = \xi(S)$.

Here $\widehat{\dim}$ denotes the maximal topological dimension over all locally compact subsets of $\mathcal{M}^\omega(S)$.

A direct consequence of this is a proof of the "rank conjecture" of Brock–Farb [30], which states that $\xi(S)$ is the maximal rank of a quasi-isometrically embedded flat in $\mathcal{M}^\omega(S)$. Independently, Hamenstädt [58] proved this theorem by somewhat different means, establishing in particular a homological version of the dimension theorem. The main idea of our proof is to study certain families of separating sets, which we describe in more detail below.

Separation properties of the asymptotic cone. In Behrstock [7] it was shown that every point of $\mathcal{M}^\omega(S)$ is a cut point. An extension and refinement of Behrstock's construction leads to the following statement. First, let $r(W) = \xi(W)$ if W is a connected non-annular essential subsurface of S, $r(W) = 1$ if W is an essential annulus, and define it over disjoint unions to be additive. It is not hard to see that Theorem 4.2 for essential (but not necessarily connected) subsurfaces W in S should state that $\widehat{\dim}\, \mathcal{M}^\omega(W) = r(W)$, where $\mathcal{M}^\omega(W)$ is the product of $\mathcal{M}^\omega(W_i)$ over the components. (When W is an annulus in particular, note that $\mathcal{M}^\omega(W)$ is $\mathcal{A}^\omega(W) = \mathbb{R}$, hence the definition $r = 1$ in that case).

Theorem 4.3. *There is a family $\mathcal{L}$ of closed subsets of $\mathcal{M}^\omega(S)$ with the following properties:*

- *Each $L \in \mathcal{L}$ is either a single point, or is bilipschitz equivalent to $\mathcal{M}^\omega(W)$ for a (possibly disconnected) subsurface $W \subset S$ with $r(W) < r(S)$.*
- *For any $x, y \in \mathcal{M}^\omega(S)$ there exists $L \in \mathcal{L}$ which separates x from y.*

We call $\mathcal{L}$ "separators" of $\mathcal{M}^\omega(S)$. The case that $L \in \mathcal{L}$ is a single point is Behrstock's cut point theorem, and the theorem on topological dimension can be obtained

by induction (in a locally compact space, the existence of a family of separators like this with dimensions at most $n-1$ implies a dimension upper bound of n for the whole space. The lower bound is easy).

The idea of the proof is to find "rank 1 directions" in the cone, and establish the separators as "transverse sets" for these directions. Let us consider the special case of $S = S_{0,5}$, let $\boldsymbol{\delta} = (\delta_n)$ be a sequence of simple closed curves in S, and let $\mathcal{Q}^\omega(\boldsymbol{\delta})$ be the ω-limit of $Q_n(\delta_n)$. We know from Theorem 4.1 that $\mathcal{Q}^\omega(\boldsymbol{\delta})$ can be identified with $\mathbb{R} \times T$, where $\mathbb{R}$ is $\mathcal{M}^\omega(\delta)$, the asymptotic cone of the sequence of twist complexes $\mathcal{A}(\delta_n)$, and T is $\mathcal{M}^\omega(S \setminus \boldsymbol{\delta})$, the asymptotic cone of $\mathcal{M}(S \setminus \delta_n)$, which is an $\mathbb{R}$-tree since $S \setminus \delta_n \cong S_{0,4}$.

A separator associated to this picture is a subset of $\mathcal{Q}^\omega(\boldsymbol{\delta})$ of the form $\{s\} \times T$, which certainly separates $\mathcal{Q}^\omega(\boldsymbol{\delta})$, and has $\widehat{\dim} = 1$ since it is an $\mathbb{R}$-tree. To show that it has global separation properties we consider the map

$$\pi_{\boldsymbol{\delta}} \colon \mathcal{M}^\omega(S) \to \mathcal{M}^\omega(\boldsymbol{\delta}) = \mathbb{R}$$

which is the rescaled ω-limit of the projection maps from $\mathcal{M}(S)$ to $\mathcal{A}(\delta_n)$. This is certainly a Lipschitz map, and restricted to $\mathcal{Q}^\omega(\boldsymbol{\delta})$ becomes projection to the first factor. Globally, it has the following useful property:

Lemma 4.4. *$\pi_{\boldsymbol{\delta}}$ is locally constant in the complement of $\mathcal{Q}^\omega(\boldsymbol{\delta})$.*

To understand why this holds, consider a point $\boldsymbol{x} \in \mathcal{M}^\omega(S) \setminus \mathcal{Q}^\omega(\boldsymbol{\delta})$. $\boldsymbol{x}$ is a sequence (or rather an equivalence class of sequences) of markings (x_n) whose distance from $Q(\delta_n)$ is growing linearly (with respect to the scale constants s_n). Using the distance formula (Theorem 3.3), one can show that $d(x_n, Q(\delta_n))$ is estimated by

$$\sum_{Y \cap \delta_n \neq \emptyset} \{d_{\mathcal{A}(Y)}(x_n, \delta_n)\}_K \tag{4}$$

which is therefore growing linearly as well. The domains in the sum (4) are partially ordered by their appearance in any hierarchy from δ_n to x_n (see Theorem 3.2). Now if y_n is given where $d_{\mathcal{M}(S)}(x_n, y_n)$ is a (sufficiently small) fraction of s_n, then the stability property in Theorem 3.2 implies that most of these domains are forced to appear in the hierarchy from δ_n to y_n as well.

In particular, the boundaries of these domains cross δ_n, and with this one can then show that $\pi_{\delta_n}(x_n)$ and $\pi_{\delta_n}(y_n)$ are a bounded distance apart. In the rescaling limit, we conclude that there is a neighborhood of $\boldsymbol{x}$ on which $\pi_{\boldsymbol{\delta}}$ is constant.

Once this lemma is established, we see immediately that the complement of $L_s = \{s\} \times T$ in $\mathcal{M}^\omega(S)$ is a union of three disjoint open sets: $\pi_{\boldsymbol{\delta}}^{-1}((-\infty, s))$, $\pi_{\boldsymbol{\delta}}^{-1}((s, \infty))$, and $\pi_{\boldsymbol{\delta}}^{-1}(s) \setminus L_s$. This gives the desired separation property.

In the general setting, the tricky question is what should take the place of the $\mathbb{R}$ factor in $\mathcal{Q}^\omega(\boldsymbol{\delta})$. One might for example take a sequence $\boldsymbol{W} = (W_n)$ and look at the product decomposition

$$\mathcal{Q}^\omega(\partial \boldsymbol{W}) \cong \mathcal{M}^\omega(\boldsymbol{W}) \times \mathcal{M}^\omega(\boldsymbol{W}^c)$$

(where $\boldsymbol{W}^c = (W_n^c)$, and W_n^c denotes all the pieces except W_n in the partition $\sigma(\partial W_n)$.) There is a projection $\mathcal{M}^\omega(S) \to \mathcal{M}^\omega(\boldsymbol{W})$, but it does *not* have the required locally constant properties.

Instead we identify a certain decomposition of $\mathcal{M}^\omega(\boldsymbol{W})$ into $\mathbb{R}$-trees, and for each such tree F we consider the set

$$\mathcal{P}_F = F \times \mathcal{M}^\omega(\boldsymbol{W}^c) \subset \mathcal{Q}^\omega(\partial \boldsymbol{W}).$$

F now plays the role of the $\mathbb{R}$ factor in the example – there is a map $\mathcal{M}^\omega(S) \to F$ which is locally constant outside $\mathcal{P}_F$, and is projection to the first factor within $\mathcal{P}_F$. The sets $\{s\} \times \mathcal{M}^\omega(\boldsymbol{W}^c)$ are our separators.

This family of product regions and retractions gives a tractable structure for analyzing $\mathcal{M}^\omega(S)$. There is a reasonable hope that these techniques can lead to a good understanding of bilipschitz flats in $\mathcal{M}^\omega(S)$ and hence quasiflats in $\mathcal{M}(S)$, and more generally to a global understanding of the topology of the asymptotic cone. In particular we are hopeful, at the time of this writing, that this should yield another approach to showing the *quasi-isometric rigidity of* $\mathcal{MCG}(S)$. (This is the property that the group of quasi-isometries of $\mathcal{MCG}(S)$ is, up to bounded error, the same as the left-action of $\mathcal{MCG}(S)$ on itself). We note that Hamenstädt has announced a proof of quasi-isometric rigidity using her technique of analyzing train-track splitting sequences, which has a somewhat different flavor.

5. Hyperbolic geometry and ending laminations

The complex of curves, and the machinery of hierarchies, play an important role in the solution of Thurston's Ending Lamination Conjecture, a classification conjecture for the deformation space of a Kleinian group. We will give a brief description of this conjecture and its connection to the complex of curves; for more details see the expository article [102].

If G is a torsion-free, finitely generated group we may consider the set $AH(G)$ of (conjugacy classes of) discrete, faithful representations of G into $\mathrm{PSL}_2(\mathbb{C})$, the isometry group of hyperbolic 3-space $\mathbb{H}^3$. Equivalently $AH(G)$ is the set of marked hyperbolic 3-manifolds with fundamental group G. Let $N_\rho = \mathbb{H}^3/\rho(G)$ for $\rho \in AH(G)$.

Mostow/Prasad rigidity states that, if $AH(G)$ contains an element with finite volume, then in fact $AH(G)$ is a singleton – there is a unique hyperbolic manifold in the corresponding homotopy class. In the infinite volume case, there is a rich deformation theory, in which the dominant theme is that the geometry of a hyperbolic 3-manifold is controlled by its *ends*, which are in turn described by deformation theory of surfaces.

Let us sketch this structure briefly in rough historical order, focusing first on the case that G is $\pi_1(S)$, for a closed surface S (more about the general case below).

Fuchsian groups. If ρ takes values in $\mathrm{PSL}_2(\mathbb{R})$ then it leaves invariant the circle $\mathbb{R} \cup \infty$ in the Riemann sphere, and its convex hull the hyperbolic plane $\mathbb{H}^2 \subset \mathbb{H}^3$. The quotient is diffeomorphic to $S \times \mathbb{R}$, with a well-known warped product metric.

Kleinian groups. Poincaré perturbed Fuchsian groups in $\mathrm{PSL}_2(\mathbb{C})$ to obtain groups that are still topologically (in fact quasiconformally) conjugate to Fuchsian groups. In particular their quotients are still $S \times \mathbb{R}$. The circle becomes a Jordan curve, now defined as the *limit set* of ρ.

QC deformation theory. Bers and Ahlfors [2], [9] showed how to parameterize the entire quasiconformal deformation space of a Fuchsian group as $\mathcal{T}(S) \times \mathcal{T}(S)$. The idea is that the two sides of the limit set, modulo the group, give two Riemann surfaces, and conversely (the hard part) these Riemann surfaces can be arbitrarily prescribed using the Measureable Riemann Mapping Theorem.

Degenerations. Bers, Greenberg, Kra, Maskit, Marden and Sullivan [10], [12], [52, 83], [91], [89], [87], [88], [127] studied spaces of quasiconformal deformations, and particularly limits in which cusps are formed, and more exotic "degenerate groups" occur. The action on the Riemann sphere is no longer conjugate to the Fuchsian action (in particular the limit set is no longer a Jordan curve) but the quotient of $\mathbb{H}^3$ is still $S \times \mathbb{R}$ (this was initially known only in the case where the degeneration only involved cusp formation, i.e. the *geometrically finite* case).

Geometric tameness. Thurston proposed the notion of a geometrically tame group, whose convex core is swept out by pleated surfaces (see § 1). The convex core is the quotient of the hyperbolic convex hull of the limit set, and is also the smallest convex submanifold of N_ρ carrying the fundamental group. Geometrically finite manifolds have convex hulls of finite volume. He showed that if ρ is geometrically tame then N_ρ is still homeomorphic to $S \times \mathbb{R}$, and established tameness for certain limits of quasiconformal deformations. He defined the notion of *ending laminations* which, for the degenerate ends, take the place of the Riemann surfaces of Ahlfors–Bers.

Tameness for surface groups. Bonahon [19] established tameness for *all* surface groups $\rho \in AH(\pi_1(S))$, and more generally for all groups that have no free-product decomposition.

So far the description of $\rho \in AH(\pi_1(S))$ is that N_ρ is diffeomorphic to $S \times \mathbb{R}$, has two ends (corresponding to the two ends of $\mathbb{R}$), each of which has either a "Riemann surface at infinity" which describes its asymptotic structure, or an "ending lamination" which we describe now. (We are oversimplifying – when cusps are present we should cut along them and obtain a larger collection of ends).

In language amenable to our curve-complex discussion, we can define the ending laminations as follows: If one end of N_ρ is degenerate, it is filled by a sequence of pleated surfaces homotopic to $S \times \{0\}$. Each such surface contains simple curves of uniformly bounded length, and these must go to infinity in $\mathcal{C}(S)$. They converge to a point in $\partial\mathcal{C}(S) = \mathcal{EL}(S)$ (this is by no means obvious and is a major ingredient in the work of both Thurston and Bonahon).

Thurston's Ending Lamination Conjecture is then the following statement:

Theorem 5.1 (Brock–Canary–Minsky [99], [29], [28]). *A hyperbolic* 3*-manifold with finitely-generated fundamental group is uniquely determined by its topological type and its list of end invariants.*

This theorem gives a complete classification of the deformation space of a Kleinian group, and this has a number of consequences. For example, it is an ingredient of the proof of the Bers–Sullivan–Thurston Density Conjecture, which states that the geometrically finite (or equivalently structurally stable) representations are dense in the deformation space of any group. (This proof, whose outline was in place since the work of Thurston and Bonahon, required for its completion also the work of Kleineidam–Souto [80], Lecuire [84], Kim–Lecuire–Ohshika [71], and Namazi–Souto. An alternate treatment was given by Rees [114]. A completely different, rather surprising proof which works at least in the incompressible-boundary case was given by Bromberg [33] and Brock–Bromberg [27], before Theorem 5.1 was established.)

On the other hand, the Ending Lamination Conjecture does not give us a complete understanding of the geometry and topology of Kleinian deformation spaces. In fact this structure turns out to be considerably more complex than originally suspected, and perhaps too complex for any neat description. See e.g. Anderson–Canary–McCullough [3], Bromberg [34], Bromberg–Holt [36], Ito [73], McMullen [97].

A little about the proof. We begin by attempting to understand the *bounded-length curves* in N_ρ – that is to locate the homotopy classes in S of those curves whose geodesic representatives in N_ρ satisfy some fixed length bound. As above, many such curves exist because of the presence of pleated surfaces.

The structure of $\mathcal{C}(S)$ comes in naturally here because, as observed in the introduction, if $[vw]$ is an edge in $\mathcal{C}(S)$ then there is a pleated surface mapping v and w simultaneously to geodesics. Thus we study the *length function*

$$\ell_\rho \colon v \mapsto \ell(\rho(v))$$

on $\mathcal{C}(S)$, associating to each vertex the length of the corresponding closed geodesic in N_ρ. The ending laminations are accumulation points, on the boundary at infinity $\partial\mathcal{C}(S) \cong \mathcal{EL}(S)$, of the set $\{\ell_\rho \le L_0\}$ where L_0 is a fixed constant depending on the topology of S. The initial geometric result linking the geometry of N_ρ with that of $\mathcal{C}(S)$ is

Theorem 5.2 ([100]). *The sublevel set*

$$\{v \in \mathcal{C}(S) : \ell_\rho(v) \le L\}$$

is quasiconvex for $L \ge L_0$.

This gives some kind of very rough control, but as in the discussion of the mapping class group, we need to somehow refine this by looking at the link structure of

the complex. This is accomplished in [99], where we show that, for any essential subsurface $W \subset S$, the projection

$$\pi_{\mathcal{A}(W)}(\{v : \ell_\rho(v) \leq L\}) \tag{5}$$

is also quasiconvex. In the end, this refined control leads to a theorem about hierarchies. We extend the notion of a finite hierarchy to an infinite one whose "endpoints" are filling laminations, and establish a statement of this type:

Theorem 5.3 ([99]). *Given an end-invariant pair (ν_-, ν_+) in a surface S, there exists a hierarchy of geodesics H in $\mathcal{C}(S)$ connecting ν_- to ν_+, such that, if $\rho \in AH(\pi_1(S))$ has end invariants $\nu_\pm$ then*

1. *All the vertices that appear in H have uniformly bounded ρ-length.*
2. *All sufficiently ρ-short curves do appear in H, and the geometry of their Margulis tubes can be estimated from the data in H.*
3. *The order in which the Margulis tubes are arranged in N_ρ is consistent with the order of hierarchy paths of H.*

By "sufficiently" and "uniformly", we refer to bounds that depend only on the topological type of S, and not on ρ or $(\nu_\pm)$.

In order to prove this theorem, we study the map $\Pi\colon \mathcal{M}(S) \to \mathcal{M}(S)$ which maps a marking μ to a marking of minimal length in a pleated surface that maps base(μ) to its geodesic representative in N_ρ. In fact this map plays an important role in the proof of Theorem 5.2 and its generalization (5). Its "shadow" in the complex of curves, for example, is a map from $\mathcal{C}(S)$ to $\{v \in \mathcal{C}(S) : \ell_\rho(v) \leq L\}$ which we show is coarsely Lipschitz and coarsely the identity on its target – i.e. a coarse Lipschitz retraction. In a δ-hyperbolic space, the image of such a map is quasi-convex, and furthermore has the stability property that any geodesic with endpoints in $\{\ell_\rho(v) \leq L\}$ must be mapped uniformly close to itself.

A generalization of this argument shows, for any μ in a hierarchy path associated to H, that

$$d_{\mathcal{A}(W)}(\mu, \Pi(\mu))$$

is uniformly bounded, for *any* essential W. The distance formula of Theorem 3.3 now shows that μ and $\Pi(\mu)$ are a uniform distance apart in $\mathcal{M}(S)$, and in particular it follows that the length of base(μ) in N_ρ is uniformly bounded. These *a priori* bounds are the main step, and make the proof of the rest of the theorem possible.

Theorem 5.3 enables us to build a *combinatorial model* M_ν for the geometry of N_ρ, which depends only on the end invariants themselves, and show (in [29]) that M_ν and N_ρ are bilipschitz homeomorphic. Hence two N_ρ's with the same end invariants are bilipschitz equivalent to each other, and Sullivan's rigidity theorem [126] then implies that they are isometric.

The structure of the model, very roughly, is this: M_ν, which we identify topologically with $S \times \mathbb{R}$, contains a union $\mathcal{U}$ of level solid tori, i.e. regular neighborhoods of

curves of the form $\gamma \times \{t\}$. Each of these solid tori is given the geometry of a Margulis tube. The complement of $\mathcal{U}$ is broken up into "blocks", each of which falls into a finite number of topological types $F \times [-1, 1]$ and a compact set of geometric shapes. The tubes and the boundaries of the block surfaces F all correspond to vertices in the hierarchy. (We are oversimplifying a bit by not describing the cases where there are cusps, or where the convex core is not the whole manifold). See [101] for a detailed exposition of this construction when $S = S_{0,5}$.

The general case. Most of the geometry discussed above applies to any π_1-injectively immersed surface in a hyperbolic 3-manifold, by considering the appropriate cover. In particular, a general hyperbolic N with finitely-generated fundamental group has a *compact core* $K \subset N$ (Scott [121], McCullough [96]) whose inclusion is a homotopy-equivalence, and when ∂K is incompressible we can apply the same technique to understand the *ends* of N, i.e. the components of $N \setminus K$.

When K has compressible boundary the situation is more delicate, but a collection of, by now, well-known techniques, together with the resolution by Agol [1] and Calegari–Gabai [39] of Marden's Tameness Conjecture, allow us again to extend the arguments.

The Tameness Conjecture (now theorem) states that N is homeomorphic to the interior of K, or equivalently K can be chosen so that $N \setminus K \cong \partial K \times \mathbb{R}$. (This was not obvious in the incompressible boundary case either, but was established by the work of Thurston and Bonahon [130], [19]).

Canary [40] showed that the end of a tame hyperbolic 3-manifold is isometric to an *incompressible* end of a 3-manifold with *pinched negative curvature*. This is done by means of lifting to an appropriately chosen branched cover. Working in this branched cover we can apply the techniques that worked in the incompressible boundary case to obtain again a bilipschitz model for the ends.

One disadvantage of this technique is that *uniformity* of the model no longer holds. Whereas in the case of $S \times \mathbb{R}$ the quality of our estimates depended only on the topological type of S, in general even the topology of the branched cover construction depends on the geometry, via the choice of branching locus. Outside some compact set our model does have uniform quality, but there is no uniform control on the size of this compact set, and hence no uniform overall model. This issue plays a role in the next section as well.

Our proof in the compressible-boundary case [28] has still, alas, not appeared. However, accounts (with alternative proofs) are available by Bowditch [23] and Rees [114]. Namazi [106] has also written down some of the tools needed to attack the compressible case.

6. Heegaard splittings

By Mostow's theorem, the geometry of a closed hyperbolic 3-manifold is uniquely determined by its topology. There does not, however, presently exist an effective general way of describing the geometry from topological data. Let us examine a specific version of this question, namely how the data of a Heegaard splitting give us geometric (and for that matter, topological) information.

The genus, of course, is the first interesting piece of data, and in particular Heegaard splittings of minimal genus (called the *Heegaard genus*) should be of interest. The genus clearly bounds from above the rank of the fundamental group, but no general bound exists in the opposite direction. Boileau–Zieschang [18] first gave examples with rank 2 and genus 3. Irreducible examples giving arbitrarily large difference between genus and rank have been found by Schultens–Weidmann [120] – these examples are all graph manifolds. There are no hyperbolic examples with rank smaller than genus, and one may speculate that they are equal, or more conservatively that rank gives some explicit upper bound on genus in the hyperbolic setting. Some partial results in this direction are due to Namazi-Souto [107] and Souto [125]. Bachman–Cooper–White [4] have shown that the Heegaard genus gives an upper bound for the injectivity radius at any point in a hyperbolic 3-manifold. Their methods involve "sweepouts" a la Pitts–Rubinstein, and should have further applications (see also Souto [125]). There are many related questions, e.g. behavior of tunnel number for knot complements, which we will not touch here. Instead we will concentrate on fixing a genus and obtaining geometric information from the complexity of the gluing map between the two handlebodies.

Let H_+ and H_- denote the pair of handlebodies of a Heegard splitting of a 3-manifold M, with $S = \partial H_+ = \partial H_-$. This defines two *meridian sets*, $\mathcal{D}_+$ and $\mathcal{D}_-$ in $\mathcal{C}(S)$, namely those simple curves that are the boundaries of essential disks in H_+ and H_- respectively. In turn, $\mathcal{D}_\pm$ determine the pair (M, S) uniquely. Note that $\mathcal{D}_\pm$ contain all possible Heegaard diagrams for the splitting.

Hempel distance. Hempel [68] pointed out that the quantity

$$d(M, S) = d_{\mathcal{C}(S)}(\mathcal{D}_+, \mathcal{D}_-),$$

called "Heegaard distance" or "Hempel distance" (where $d_{\mathcal{C}(S)}$ denotes minimal distance between the two sets), is a useful indication of complexity which generalizes the notions of weak and strong reducibility we mentioned in the introduction. Indeed it is not hard to see that the definitions from the introduction translate this way:

- $d(M, S) = 0$ iff the splitting is reducible;
- $d(M, S) \leq 1$ iff the splitting is weakly reducible;
- $d(M, S) \leq 2$ iff the splitting has the disjoint curve property;
- $d(M, S) \geq 2$ iff the splitting is strongly irreducible.

Moreover Hempel showed that there are splittings with arbitrarily large distance. As mentioned in the introduction (see also Thompson [129]), Hempel showed that if M is Seifert fibred or has an essential torus then $d(M, S) \leq 2$. One can also check that $d(M, S) \leq 2$ for all remaining geometrizable non-hyperbolic cases, and Hempel therefore conjectured that

$$d(M, S) \geq 3 \quad \Longrightarrow \quad M \text{ is hyperbolic.}$$

Perelman's work on Thurston's geometrization conjecture, once accepted, implies that this conjecture holds.

Schleimer [119] proved that, if M is fixed, then there are at most finitely many (isotopy classes of) Heegaard surfaces with $d(M, S) > 2$. This is a sort of rigidity property for high-distance splittings, and suggests that they should convey topological information.

Hartshorn [62] obtained upper bounds on Hempel distance in the presence of incompressible surfaces, using ideas of Kobayashi [81]. Bachman–Schleimer [5], [6] obtain related results for knots and surface bundles. Scharlemann–Tomova [118] have obtained distance bounds from the interaction of pairs of splittings in a manifold.

However, one soon observes that $d(M, S)$ by itself is far from being an accurate measure of complexity. For example, for fixed g one can find hyperbolic manifolds, of arbitrarily high volume, which have a (minimal) genus g splitting of distance 1 (see Souto [123]).

One should really look at a more refined comparison of $\mathcal{D}_+$ and $\mathcal{D}_-$, much as in § 5 we looked at hierarchies connecting end invariants rather than plain geodesics in $\mathcal{C}(S)$ connecting them.

If $\mathcal{P}_+$ is the set of pants decompositions of S composed of meridian curves in H_+ (and similarly $\mathcal{P}_-$) we could for example consider the combinatorial distance $d(\mathcal{P}_+, \mathcal{P}_-)$, by which we mean the distance in the "graph of pants decompositions" as in [26]. This distance, by a triangulation argument, gives an upper bound for the hyperbolic volume. Brock–Souto [31] have formulated a slightly different combinatorial distance (enlarging $\mathcal{P}_\pm$ to include more pants decompositions) which gives both upper and lower bounds.

Geometric models for handlebodies. One is naturally led from Heegaard splittings into the related problem of describing hyperbolic structures on a *single* handlebody. If one were to have a recipe for handlebodies one could attempt to glue them together and obtain models for Heegaard splittings. Namazi carried this out in [106] for a special family of splittings.

A partial answer was provided by the Ending Lamination construction, namely a geometric model for the end of the manifold. As mentioned above, this model was lacking in uniformity. In other words the work of [28] partitions a hyperbolic handlebody N into a compact handlebody K and an end E homeomorphic to $\partial K \times (0, \infty)$.

We provide a bilipschitz model for E of uniform quality, but give no information at all about the geometry of K.

What sort of geometric pictures do we expect to find? Let us describe a motivating list of examples, drawn from work of Namazi and Brock–Souto. When we say a subset of a manifold has *bounded geometry* here, we mean that, within the family of examples in question, it is drawn from a compact set of possibilities.

Capped-off products. One can construct hyperbolic handlebodies H for which there is a decomposition $H = E \cup K$, with K a compact handlebody with bounded geometry and E bilipschitz-equivalent to an end of an $S \times \mathbb{R}$. Indeed, Namazi [106] exhibited a family of such manifolds whose end invariants satisfy a *bounded combinatorics* condition, and which furthermore satisfy a lower bound on injectivity radius.

Small 1*-handles.* There are hyperbolic handlebodies H which admit a decomposition $H = E \cup B \cup K_1 \cup K_2$, where K_1 and K_2 are handlebodies of lower genus, each K_i is bilipschitz equivalent to the convex hull of a "capped off product" as above, B is a 1-handle of bounded geometry joining K_1 to K_2, and E is bilipschitz equivalent to an end of an $S \times \mathbb{R}$, and is attached along its bottom boundary to the handlebody $K_1 \cup B \cup K_2$.

Brock–Souto show that this (with variations in the number and arrangement of the pieces and 1-handles) is the structure of a general hyperbolic handlebody with a uniform lower bound on injectivity radius

Incompressible I-bundles. After representing a (closed) handlebody as $F \times [-1, 1]$ where F is a compact surface with boundary, let γ be a copy of $\partial F \times \{0\}$ pushed slightly into the interior H. One can find a hyperbolic structure on $H \setminus \gamma$ which makes the components of γ into cusps, and then after a geometric step known as "Dehn filling" obtain a hyperbolic structure on H where γ is extremely short. This can be done so that the geometry of H, outside the Margulis tubes of γ, is essentially that of a surface group with cusps based on F, and hence is described by a model manifold of the type used for the Ending Lamination Conjecture.

Tubes with 1*-handles* One can start with a finite number of Margulis tubes (solid tori), and join them with 1-handles of bounded geometry.

These examples can be combined – for example the tubes with 1-handles can serve as a core to which a standard product end is added, and the product structure example can be attached to something else using a 1-handle, or a peripheral annulus. This can be done geometrically via the Klein–Maskit combination theorem [90], the Ahlfors–Bers deformation theory [11], and variations on Thurston's Dehn filling construction (see Bonahon–Otal [21], Bromberg [32], [35], Comar [45], Hodgson–Kerckhoff [70]).

We expect that any handlebody will have a decomposition into pieces of this type. An interesting challenge is to *predict* from the topological data what pieces actually occur. In other words, starting with a pair $(\mathcal{D}, \mu)$ where $\mathcal{D}$ is a meridian set and μ an end invariant on S, we would like to give such a decomposition, whose structure and the shape of the pieces (e.g. boundary structure of the Margulis tubes, and end invariants of the I-bundle pieces) can be read off from the input data.

Geometry of $\mathcal{D}$. A first step toward the goal of providing uniform geometric models is to understand the structure of $\mathcal{D}$ and how it embeds in $\mathcal{C}(S)$. With Masur in [92] we proved

Theorem 6.1. *$\mathcal{D}$ is a quasi-convex subset of $\mathcal{C}(S)$.*

This is analogous to Theorem 5.2 on quasiconvexity of the bounded-curve set in the incompressible setting.

The next step might be to analyze projections of $\mathcal{D}$ into subsurfaces. Very roughly, one expects surfaces $W \subset S$ in which $d_W(\mathcal{D}, \mu)$ is very large to play a role in the hyperbolic geometry associated to the end invariant μ – much as subsurfaces where $d_W(\nu_+, \nu_+)$ is large played a role in the geometry of surface groups with end invariants ν_+ and ν_-. For the purpose of understanding when this happens, those W for which $\operatorname{diam}_W(\mathcal{D})$ is bounded are somewhat easier to analyze. Masur–Schleimer have studied the structure of $\mathcal{D}$ quite extensively in [93], and have shown in particular that

Theorem 6.2 (Masur–Schleimer). *Let S be the boundary of a handlebody H. If W is an essential subsurface of S then $\operatorname{diam}_W(\mathcal{D})$ is bounded by a number depending on the genus of S, unless*

1. *there is a meridian in the complement of W,*
2. *there is a meridian in W but not in its complement,*
3. *there is an essential I-bundle B in H with W a component of its horizontal boundary, and at least one vertical annulus of B lying in S.*

In cases 1 *and* 3, *$\pi_W(\mathcal{D})$ is all of $\mathcal{A}(W)$ up to bounded gaps. In case* 2, *$\pi_W(\mathcal{D})$ is within a bounded neighborhood of $\mathcal{D} \cap \mathcal{C}(W)$.*

This is part of a larger analysis in which, in particular, they show that the Hempel distance of a splitting can be estimated algorithmically.

There is an interplay between this classification of subsurfaces and the geometric examples we listed above. To understand case 3, for example, note that in a product $F \times [-1, 1]$, where F is a surface with boundary, there are essential disks of the form $a \times [-1, 1]$ where a is any essential arc in F. Hence (supposing $F \times [-1, 1]$ to be embedded in the handlebody H as in part (3)) we see that $\pi_{F\times\{1\}}(\mathcal{D})$ gives all arcs of $\mathcal{A}(F)$. This corresponds to the "incompressible I-bundle" case in the list of geometric examples, and we expect in this case that the projections of μ to $F \times \{\pm 1\}$ should act as end invariants for this incompressible surface group within H.

With these results as a starting point, one can hope that there is a model based, in a way analogous to hierarchies, on the shortest path from a marking μ to the meridian set $\mathcal{D}$. The parts of the path far from $\mathcal{D}$ should give a portion of the model analogous to the compressible products in the example list, whereas near $\mathcal{D}$ the construction would need to give rise to a system of 1-handles and incompressible I-bundles. A successful version of this would give us uniform bilipschitz models for handlebodies,

and via some deformation and gluing could enable us to build models for closed hyperbolic 3-manifolds as well. One would also like to apply our understanding of $\mathcal{D}$ directly to the combinatorics of a Heegaard splitting, that is to the relative positioning of a pair $(\mathcal{D}_+, \mathcal{D}_-)$.

References

[1] Agol, I., Tameness of hyperbolic 3-manifolds. Preprint, 2004; arXiv:math.GT/0405568.

[2] Ahlfors, L., and Bers, L., Riemann's mapping theorem for variable metrics. *Ann. of Math.* **72** (1960), 385–404.

[3] Anderson, J., Canary, R., and McCullough, D., The topology of deformation spaces of Kleinian groups. *Ann. of Math.* (2) **152** (3) (2000), 693–741.

[4] Bachman, D., Cooper, D., and White, M. E., Large embedded balls and Heegaard genus in negative curvature. *Algebr. Geom. Topol.* **4** (2004), 31–47.

[5] Bachman, D., and Schleimer, S., Distance and bridge position. arXiv:math.GT/0308297.

[6] —, Surface bundles versus Heegaard splittings. arXiv:math.GT/0212104.

[7] Behrstock, J., Asymptotic geometry of the Mapping Class Group and Teichmüller space. Ph.D. thesis, SUNY at Stony Brook, 2004.

[8] Behrstock, J., and Minsky, Y., Dimension and rank for mapping class groups. 2005, arXiv:math.GT/0512352.

[9] Bers, L., Simultaneous uniformization. *Bull. Amer. Math. Soc.* **66** (1960), 94–97.

[10] —, On boundaries of Teichmüller spaces and on Kleinian groups I. *Ann. of Math.* **91** (1970), 570–600.

[11] —, Spaces of Kleinian groups. In *Several Complex Variables, I* (University of Maryland, College Park, Md., 1970), Lecture Notes in Math. 155, Springer-Verlag, Berlin 1970, 9–34.

[12] —, Spaces of degenerating Riemann surfaces. In *Discontinuous groups and Riemann surfaces*, Ann. of Math. Stud. 79, Princeton University Press, Princeton, N.J., 1974, 43–59.

[13] Birman, J., Lubotzky, A., and McCarthy, J., Abelian and solvable subgroups of the mapping class groups. *Duke Math J.* **50** (1983), 1107–1120.

[14] Joan S. Birman, The topology of 3-manifolds, Heegaard distance and the mapping class group of a 2-manifold. arXiv:math.GT/0502545.

[15] Boileau, M., Collins, D. J., and Zieschang, H., Genus 2 Heegaard decompositions of small Seifert manifolds. *Ann. Inst. Fourier* (*Grenoble*) **41** (4) (1991), 1005–1024.

[16] Boileau, M., and Otal, J.-P., Sur les scindements de Heegaard du tore T^3. *J. Differential Geom.* **32** (1) (1990), 209–233.

[17] Boileau, M., Rost, M., and Zieschang, H., On Heegaard decompositions of torus knot exteriors and related Seifert fibre spaces. *Math. Ann.* **279** (3) (1988), 553–581.

[18] Boileau, M., and Zieschang, H., Heegaard genus of closed orientable Seifert 3-manifolds. *Invent. Math.* **76** (3) (1984), 455–468.

[19] Bonahon, F., Bouts des variétés hyperboliques de dimension 3. *Ann. of Math.* **124** (1986), 71–158.

[20] Bonahon, F., and Otal, J.-P., Scindements de Heegaard des espaces lenticulaires. *Ann. Sci. École Norm. Sup.* (4) **16** (3) (1983), 451–466.

[21] Bonahon, F., and Otal, J.-P., Variétés hyperboliques à géodésiques arbitrairement courtes. *Bull. London Math. Soc.* **20** (1988), 255–261.

[22] Bowditch, B., Notes on Gromov's hyperbolicity criterion for path-metric spaces. In *Group theory from a geometrical viewpoint* (Trieste, 1990), World Scientific Publishing, River Edge, NJ, 1991, 64–167.

[23] —, End invariants of hyperbolic 3-manifolds. Preprint, Southampton, 2005.

[24] Bridson, M. R., Geodesics and curvature in metric simplicial complexes. In *Group theory from a geometrical viewpoint* (Trieste, 1990), World Scientific Publishing, River Edge, NJ, 1991, 373–463.

[25] Bridson, M. R., and Haefliger, A., *Metric spaces of non-positive curvature*. Grundlehren Math. Wiss. 319, Springer-Verlag, Berlin 1999.

[26] Brock, J., The Weil-Petersson metric and volumes of 3-dimensional hyperbolic convex cores. Preprint, 2001.

[27] Brock, J., and Bromberg, K., On the density of geometrically finite Kleinian groups. Preprint, 2002.

[28] Brock, J., Canary, R., and Minsky, Y., The classification of finitely-generated Kleinian groups. In preparation.

[29] —, The classification of Kleinian surface groups II: the ending lamination conjecture. arXiv:math.GT/0412006.

[30] Brock, J., and Farb, B., Curvature and rank of Teichmüller space. Preprint, 2001.

[31] Brock, J., and Souto, J., Volume and distances in the pants complex. In preparation.

[32] Bromberg, K., Hyperbolic Dehn surgery on geometrically infinite 3-manifolds. Preprint, arXiv:math.GT/000915.

[33] —, Projective structures with degenerate holonomy and the Bers density conjecture. Preprint.

[34] —, The topology of the space of punctured torus groups. In preparation.

[35] —, Rigidity of geometrically finite hyperbolic cone-manifolds. *Geom. Dedicata* **105** (2004), 143–170.

[36] Bromberg, K., and Holt, J., Self-bumping of deformation spaces of hyperbolic 3-manifolds. *J. Differential Geom.* **57** (1) (2001), 47–65.

[37] Brooks, R., and Matelski, J. P., Collars in Kleinian groups. *Duke Math. J.* **49** (1) (1982), 163–182.

[38] Buser, P., *Geometry and Spectra of Compact Riemann Surfaces*. Progr. Math. 106, Birkhäuser, Boston, MA, 1992.

[39] Calegari, D., and Gabai, D., Shrinkwrapping and the taming of hyperbolic 3-manifolds. arXiv:math.GT/0407161.

[40] Canary, R. D., Ends of hyperbolic 3-manifolds. *J. Amer. Math. Soc.* **6** (1993), 1–35.

[41] Cannon, J. W., The combinatorial structure of cocompact discrete hyperbolic groups. *Geom. Dedicata* **16** (2) (1984), 123–148.

[42] —, The theory of negatively curved spaces and groups. In *Ergodic theory, symbolic dynamics, and hyperbolic spaces* (Trieste, 1989), Oxford Sci. Publ., Oxford University Press, New York 1991, 315–369.

[43] Casson, A. J., and Bleiler, S. A., *Automorphisms of surfaces after Nielsen and Thurston*. London Math. Soc. Stud. Texts 9, Cambridge University Press, Cambridge 1988.

[44] Casson, A. J., and Gordon, C. McA., Reducing Heegaard splittings. *Topology Appl.* **27** (3) (1987), 275–283.

[45] Comar, T., Hyperbolic Dehn surgery and convergence of Kleinian groups. Ph.D. thesis, University of Michigan, 1996.

[46] Cooper, D., and Scharlemann, M., The structure of a solvmanifold's Heegaard splittings. *Turkish J. Math.* **23** (1999), 1–18.

[47] Van den Dries, L., and Wilkie, A., On Gromov's theorem concerning groups of polynomial growth and elementary logic. *J. Algebra* **89** (1984), 349–374.

[48] Farb, B., Lubotzky, A., and Minsky, Y., Rank-1 phenomena for mapping class groups. *Duke Math. J.* **106** (3) (2001), 581–597.

[49] Fathi, A., Laudenbach, F., and Poenaru, V., Travaux de Thurston sur les surfaces. *Asterisque* **66–67** (1979).

[50] Frohman, C., and Hass, J., Unstable minimal surfaces and Heegaard splittings. *Invent. Math.* **95** (3) (1989), 529–540.

[51] Gardiner, F. P., *Teichmüller theory and quadratic differentials*. Pure Appl. Math. (N.Y.), John Wiley & Sons Inc., New York 1987.

[52] Greenberg, L., Fundamental polyhedra for kleinian groups. *Ann. of Math.* (2) **84** (1966), 433–441.

[53] Gromov, M., Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73.

[54] —, Hyperbolic groups. In *Essays in Group Theory* (ed. by S. M. Gersten), Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987.

[55] —, Asymptotic invariants of infinite groups. In *Geometric group theory* (Sussex, 1991), Vol. 2, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge, 1993, 1–295.

[56] Haken, W., Some results on surfaces in 3-manifolds. Studies in Modern Topology, Math. Assoc. Amer./Prentice-Hall, Englewood Cliffs, N.J., 1968, 39–98.

[57] Hamenstädt, U., Geometry of the mapping class groups II: Subsurfaces. arXiv:math. GR/0511349.

[58] —, Geometry of the mapping class groups III: Geometric rank. arXiv:math.GT/0512429.

[59] —, Train tracks and the Gromov boundary of the complex of curves. Preprint, 2004.

[60] Harer, J., Stability of the homology of the mapping class group of an orientable surface. *Ann. of Math.* **121** (1985), 215–249.

[61] —, The virtual cohomological dimension of the mapping class group of an orientable surface. *Invent. Math.* **84** (1986), 157–176.

[62] Hartshorn, K., Heegaard splittings of Haken manifolds have bounded distance. *Pacific J. Math.* **204** (1) (2002), 61–75.

[63] Harvey, W. J., Boundary structure of the modular group. In *Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference* (ed. by I. Kra and B. Maskit), Ann. of Math. Stud. 97, Princeton University Press, Princeton, N.J., 1981.

[64] —, Modular groups and representation spaces. In *Geometry of group representations* (Boulder, CO, 1987), Contemp. Math. 74, Amer. Math. Soc., Providence, RI, 1988, 205–214.

[65] Hatcher, A., On triangulations of surfaces. *Topology Appl.* **40** (2) (1991), 189–194.

[66] —, Measured lamination spaces for surfaces, from the topological viewpoint. *Topology Appl.* **30** (1988), 63–88.

[67] Hatcher, A. E., and Thurston, W. P., A presentation for the mapping class group. *Topology* **19** (1980), 221–237.

[68] Hempel, J., *3-manifolds*. Ann. of Math. Studies 86, Princeton University Press, Princeton, N.J., 1976.

[69] —, 3-manifolds as viewed from the curve complex. *Topology* **40** (3) (2001), 631–657.

[70] Hodgson, C. D., and Kerckhoff, S. P., Universal bounds for hyperbolic Dehn surgery. *Ann. of Math.* (2) **162** (1) (2005), 367–421.

[71] Lecuire, C., Kim, I., and Ohshika, K., Convergence of freely decomposable Kleinian groups. Preprint, 2004.

[72] Imayoshi, Y., and Taniguchi, M., *An introduction to Teichmüller spaces*. Springer-Verlag, Tokyo 1992.

[73] Ito, K., Exotic projective structures and quasi-Fuchsian space. *Duke Math. J.* **105** (2) (2000), 185–209.

[74] Ivanov, N. V., *Subgroups of Teichmüller modular groups*. Transl. Math. Monogr. 115, Amer. Math. Soc., Providence, RI, 1992.

[75] —, Automorphisms of complexes of curves and of Teichmüller spaces. *Internat. Math. Res. Notices* **1997** (14) (1997), 651–666.

[76] Jørgensen, T., On discrete groups of Möbius transformations. *Amer. J. Math.* **98** (1976), 739–49.

[77] Kazhdan, D., and Margulis, G., A proof of Selberg's conjecture. *Math. USSR Sb.* **4** (1968), 147–152.

[78] Keen, L., Collars on Riemann surfaces. In *Discontinuous groups and Riemann surfaces* (Proc. Conf., Univ. Maryland 1973), Ann. of Math. Studies 79, Princeton University Press, Princeton, N.J., 1974, 263–268.

[79] Klarreich, E., The boundary at infinity of the curve complex and the relative Teichmüller space. Preprint, http://www.nasw.org/users/klarreich/research.htm.

[80] Kleineidam, G., and Souto, J., Algebraic convergence of function groups. *Comment. Math. Helv.* **77** (2) (2002), 244–269.

[81] Kobayashi, T., Heights of simple loops and pseudo-Anosov homeomorphisms. In *Braids* (Santa Cruz, CA, 1986), Contemp. Math. 78, Amer. Math. Soc., Providence, RI, 1988, 327–338.

[82] Korkmaz, M., Automorphisms of complexes of curves on punctured spheres and on punctured tori. *Topology Appl.* **95** (2) (1999), 85–111.

[83] Kra, I., On spaces of Kleinian groups. *Comment. Math. Helv.* **47** (1972), 53–69.

[84] Lecuire, C., An extension of Masur domain. In*Spaces of Kleinian Groups* (ed. by Y. Minsky, M. Sakuma, and C. Series), London Math. Soc. Lecture Note Ser. 329, Cambridge University Press, Cambridge 2006.

[85] Levitt, G., Foliations and laminations on hyperbolic surfaces. *Topology* **22** (1983), 119–135.

[86] Luo, F., Automorphisms of the complex of curves. *Topology* **39** (2) (2000), 283–298.

[87] Marden, A., The geometry of finitely generated Kleinian groups. *Ann. of Math.* **99** (1974), 383–462.

[88] Marden, A., and Maskit, B., On the isomorphism theorem for Kleinian groups. *Invent. Math.* **51** (1979), 9–14.

[89] Maskit, B., On boundaries of Teichmüller spaces and on Kleinian groups II. *Ann. of Math.* **91** (1970), 607–639.

[90] —, *Kleinian groups*. Grundlehren Math. Wiss. 287 Springer-Verlag, Berlin 1988.

[91] —, A remark on degenerate groups. *Math. Scand.* **36** (1975) (Collection of articles dedicated to Werner Fenchel on his 70th birthday), 17–20.

[92] Masur, H., and Minsky, Y., Quasiconvexity in the curve complex. In *In the Tradition of Ahlfors and Bers, III* (ed. by W. Abikoff and A. Haas), Contemp. Math. 355, Amer. Math. Soc., Providence, RI, 2004, 309–320.

[93] Masur, H., and Schleimer, S., The geometry of the disk complex. Preprint, 2006.

[94] Masur, H. A., and Minsky, Y., Geometry of the complex of curves I: Hyperbolicity. *Invent. Math.* **138** (1999), 103–149.

[95] —, Geometry of the complex of curves II: Hierarchical structure. *Geom. Funct. Anal.* **10** (2000), 902–974.

[96] McCullough, D., Compact submanifolds of 3-manifolds with boundary. *Quart. J. Math. Oxford* **37** (1986), 299–306.

[97] McMullen, C., Complex earthquakes and Teichmüller theory. *J. Amer. Math. Soc.* **11** (2) (1998), 283–320.

[98] Milnor, J., A note on curvature and fundamental group. *J. Differential Geometry* **2** (1968), 1–7.

[99] Minsky, Y., The classification of Kleinian surface groups I: models and bounds. arXiv:math.GT/0302208.

[100] —, Kleinian groups and the complex of curves. *Geom. Topol.* **4** (2000), 117–148.

[101] —, Combinatorial and geometrical aspects of hyperbolic 3-manifolds. In *Kleinian Groups and Hyperbolic 3-Manifolds* (ed. by V. Markovic Y. Komori and C. Series), London Math. Soc. Lecture Note Ser. 299, Cambridge University Press, Cambridge 2003, 3–40.

[102] —, End invariants and the classification of hyperbolic 3-manifolds. In *Current developments in mathematics, 2002*, International Press, Somerville, MA, 2003, 181–217.

[103] Moriah, Yoav, Heegaard splittings of Seifert fibered spaces. *Invent. Math.* **91** (3) (1988), 465–481.

[104] Moriah, Y., and Schultens, J., Irreducible Heegaard splittings of Seifert fibered spaces are either vertical or horizontal. *Topology* **37** (5) (1998), 1089–1112.

[105] Mosher, L., Mapping class groups are automatic. *Ann. of Math.* **142** (1995), 303–384.

[106] Namazi, H., Heegaard splittings and hyperbolic geometry. Ph.D. thesis, SUNY at Stony Brook, May 2005.

[107] Namazi, H., and Souto, J., Heegaard splittings and pseudo-anosov maps. In preparation.

[108] Otal, J.-P., Sur le nouage des geodesiques dans les varietes hyperboliques.. *C. R. Acad. Sci. Paris Sèr. I Math.* **320** (7) (1995), 847–852.

[109] —, Les géodésiques fermées d'une variété hyperbolique en tant que nœuds. In *Kleinian groups and hyperbolic 3-manifolds* (Warwick, 2001), London Math. Soc. Lecture Note Ser. 299, Cambridge University Press, Cambridge 2003, 95–104.

[110] Rafi, K., A Combinatorial Model for the Teichmuller Metric. arXiv:math.GT/0509584.

[111] —, A characterization of short curves of a Teichmüller geodesic. Preprint: math.GT/0404227.

[112] —, Hyperbolic 3-manifolds and geodesics in Teichmüller space. PhD Thesis, SUNY at Stony Brook, 2001.

[113] Rees, M., Views of parameter space: topographer and resident. *Astérisque* **288** (2003).

[114] —, The ending laminations theorem direct from Teichmüller geodesics. 2004, arxiv:math.GT/0404007.

[115] Saito, T., Scharlemann, M., and Schultens, J., Lecture notes on generalized Heegaard splittings. arXiv:math.GT/0504167.

[116] Scharlemann, M., Heegaard splittings of compact 3-manifolds. arXiv:math.GT/0007144.

[117] Scharlemann, M., and Thompson, A., Thin position for 3-manifolds. In *Geometric topology* (Haifa, 1992), Contemp. Math. 164, Amer. Math. Soc., Providence, RI, 1994, 231–238.

[118] Scharlemann, M., and Tomova, M., Alternate Heegaard genus bounds distance. UCSB 2004-38, arXiv:math.GT/0501140.

[119] Schleimer, S., The disjoint curve property. *Geom. Topol.* **8** (2004), 77–113.

[120] Schultens, J., and Weidmann, R., On the geometric and algebraic rank of graph manifolds. Preprint.

[121] Scott, G. P., Compact submanifolds of 3-manifolds. *J. London Math. Soc.* **7** (1973), 246–250.

[122] Scott, P., The geometries of 3-manifolds. *Bull. London Math. Soc.* **15** (1983), 401–487.

[123] Souto, J., The rank of the fundamental group of hyperbolic 3-manifolds fibering over the circle. Preprint, arXiv:math.GT/0503163.

[124] —, Short curves in hyperbolic manifolds are not knotted. Preprint, 2004.

[125] —, Rank and topology of hyperbolic 3-manifolds I, preprint, 2006.

[126] Sullivan, D., On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions. In *Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference*, Ann. of Math. Stud. 97, Princeton University Press, Princeton, N.J., 1981.

[127] —, Quasiconformal homeomorphisms and dynamics II: Structural stability implies hyperbolicity for Kleinian groups. *Acta Math.* **155** (1985), 243–260.

[128] Švarc, A. S., A volume invariant of coverings. *Dokl. Akad. Nauk SSSR* (*N.S.*) **105** (1955), 32–34.

[129] Thompson, A., The disjoint curve property and genus 2 manifolds. *Topology Appl.* **97** (3) (1999), 273–279.

[130] Thurston, W., *The geometry and topology of 3-manifolds*. Princeton University Lecture Notes, online at http://www.msri.org/publications/books/gt3m, 1982.

[131] —, *Three-Dimensional Geometry and Topology*. Volume 1, ed. by Silvio Levy, Princeton Math. Ser. 35, Princeton University Press, Princeton, N.J., 1997.

[132] Waldhausen, F., Heegaard-Zerlegungen der 3-Sphäre. *Topology* **7** (1968), 195–203.

[133] Zieschang, H., On Heegaard diagrams of 3-manifolds. *On the geometry of differentiable manifolds* (Rome, 1986), *Astérisque* **163–164** (7) (1988), 247–280.

Yale University, New Haven, CT 06520, U.S.A.
E-mail: yair.minsky@yale.edu

$\mathbb{A}^1$-algebraic topology

Fabien Morel

Abstract. We present some recent results in $\mathbb{A}^1$-*algebraic topology*, which means both in $\mathbb{A}^1$-homotopy theory of schemes and its relationship with algebraic geometry. This refers to the classical relationship between homotopy theory and (differential) topology. We explain several examples of "motivic" versions of classical results: the theory of the Brouwer degree, the classification of $\mathbb{A}^1$-coverings through the $\mathbb{A}^1$-fundamental group, the Hurewicz Theorem and the $\mathbb{A}^1$-homotopy of algebraic spheres, and the $\mathbb{A}^1$-homotopy classification of vector bundles. We also give some applications and perspectives.

Mathematics Subject Classification (2000). 14F05, 19E15, 55P.

Keywords. $\mathbb{A}^1$-homotopy theory, Milnor K-theory, Witt groups.

1. The Brouwer degree

Let $n \geq 1$ be an integer and let X be a pointed topological space. We shall denote by $\pi_n(X)$ the n-th homotopy group of X. A basic fact in homotopy theory is:

Theorem 1.1. *Let $n \geq 1$, $d \geq 1$ be integers and denote by S^n the n-dimensional sphere.*

1) *If $d < n$ then $\pi_d(S^n) = 0$;*

2) *If $d = n$ then $\pi_n(S^n) = \mathbb{Z}$.*

A classical proof uses the Hurewicz Theorem and the computation of the integral singular homology of the sphere. Half of this paper is devoted to explain the analogue of these results in $\mathbb{A}^1$-homotopy theory [54], [38].

For our purpose we also recall a more geometric proof of 2) inspired by the definition of Brouwer's degree. Any continuous map $S^n \to S^n$ is homotopic to a $\mathcal{C}^\infty$-differentiable map $f\colon S^n \to S^n$. By Sard's theorem, f has at least one regular value $x \in S^n$, so that $f^{-1}(x)$ is a finite set of points in S^n and for each $y \in f^{-1}(x)$, the differential $df_y\colon T_y(S^n) \to T_x(S^n)$ of f at y is an isomorphism. The "sign" $\varepsilon_y(f)$ at y is $+1$ if df_y preserves the orientation and -1 else. The integer $\delta(f) := \sum_{y \mapsto x} \varepsilon_y(f)$ is the *Brouwer degree* of f and only depends on the homotopy class of f.

Now choose a small enough open n-ball $\mathcal{B}$ around x such that $f^{-1}(\mathcal{B})$ is a disjoint union of an open n-balls $\mathcal{B}_y$ around each y's. The quotient space $S^n/(S^n - \bigcup \mathcal{B}_y)$ is

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

homeomorphic to the wedge of spheres $\vee_y S^n$ and the quotient map $S^n \to S^n/(S^n - \mathcal{B})$ is a homotopy equivalence. The induced commutative square

$$\begin{array}{ccc} S^n & \xrightarrow{\quad f \quad} & S^n \\ \downarrow & & \downarrow \wr \\ \vee_y S^n = S^n/(S^n - \bigcup \mathcal{B}_y) & \xrightarrow[\vee_y f_y]{} & S^n/(S^n - \mathcal{B}) \end{array} \tag{1.1}$$

expresses the homotopy class of f as the sum of the homotopy classes of the f_y's, each of which being the one point compactification of the differential map df_y. This proves that the degree homomorphism $\pi_n(S^n) \to \mathbb{Z}$ is injective, thus an isomorphism.

We illustrate the algebraic situation by a simple close example. Let k be a field, let $f \in k(T)$ be a rational fraction and denote still by $f\colon \mathbb{P}^1 \to \mathbb{P}^1$ the k-morphism from the projective line to itself corresponding to f. Assume, for simplicity, that f admits a regular value x in the following strong sense (which is not the generic one): x is a rational point in $\mathbb{A}^1 \subset \mathbb{P}^1$ such that f is étale over x, such that the finite étale k-scheme $f^{-1}(x)$ consists of finitely many rational points $y \in \mathbb{A}^1$ (none being ∞), and that the differentials $\frac{df}{dt}(y)$ are each units α_y. Observe that $\mathbb{P}^1 - \{x\}$ is isomorphic to the affine line $\mathbb{A}^1$ and thus the quotient morphism $\mathbb{P}^1 \to \mathbb{P}^1/\mathbb{A}^1 := T$ a "weak $\mathbb{A}^1$-equivalence". The commutative diagram (in some category of spaces over k, see below)

$$\begin{array}{ccc} \mathbb{P}^1 & \xrightarrow{\quad f \quad} & \mathbb{P}^1 \\ \downarrow & & \downarrow \wr \\ \vee_y T = \mathbb{P}^1/(\mathbb{P}^1 - f^{-1}(x)) & \xrightarrow[\vee \hat{\alpha}_y]{} & T \end{array}$$

analogous to (1.1), also expresses f, up to $\mathbb{A}^1$-weak homotopy, as the sum of the classes of the morphisms $\hat{\alpha}_y\colon T \to T$ induced by the multiplication by α_y. The idea is that in algebraic geometry the analogue of the "sign" of a unit $u \in k^\times$, or the $\mathbb{A}^1$-homotopy class of $\hat{u}$, is its class in $k^\times/(k^\times)^2$. The set $k^\times/(k^\times)^2$ should also be considered as the set of *orientations* of the affine line over k. We observe that $\hat{u}$ is $\mathbb{A}^1$-equivalent to the "1-point compactification" of the multiplication by $u\colon \mathbb{P}^1 \to \mathbb{P}^1$, $[x, y] \mapsto [ux, y]$. If $u = v^2$, the latter is $[x, y] \mapsto [ux, y] = [vx, v^{-1}y]$ which is given by the action of the matrix $\left(\begin{smallmatrix} v & 0 \\ 0 & v^{-1} \end{smallmatrix}\right)$ of $\mathrm{SL}_2(k)$ and thus, being a product of elementary matrices, is $\mathbb{A}^1$-homotopic to the identity.

Using the same procedure as in topology, we have "expressed" the $\mathbb{A}^1$-homotopy class of f as a sum of units modulo the squares and the Brouwer degree of a morphism $\mathbb{P}^1 \to \mathbb{P}^1$ in the $\mathbb{A}^1$-homotopy category $\mathsf{H}(k)$ over k should have this flavor. Denote by $GW(k)$ the Grothendieck–Witt ring of non-degenerate symmetric bilinear forms over k, that is to say the group completion of the monoid – for the direct sum – of isomorphism classes of such forms over k, see [27]. It is a quotient of the free abelian

group on units $k^\times$. We will find that the algebraic Brouwer degree over k takes it values in $GW(k)$ by constructing for $n \geq 2$ an isomorphism

$$\mathrm{Hom}_{\mathsf{H}(k)}((\mathbb{P}^1)^{\wedge n}, (\mathbb{P}^1)^{\wedge n}) \cong \mathrm{Hom}_{\mathsf{H}_\bullet(k)}((\mathbb{P}^1)^{\wedge n}, (\mathbb{P}^1)^{\wedge n}) \cong GW(k)$$

where $\mathsf{H}_\bullet(k)$ is the pointed $\mathbb{A}^1$-homotopy category over k and $\wedge$ denotes the smash-product [54], [38]. For $n = 1$ the epimorphism $\mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\mathbb{P}^1, \mathbb{P}^1) \to GW(k)$ has a kernel isomorphic to the subgroup of squares $(k^\times)^2$.

The ring $GW(k)$ is actually the cartesian product of $\mathbb{Z}$ and $W(k)$ (the Witt ring of isomorphism classes of anisotropic forms) over $\mathbb{Z}/2$, fitting into the cartesian square

$$\begin{array}{ccc} GW(k) & \longrightarrow & \mathbb{Z} \\ \downarrow & & \downarrow \\ W(k) & \longrightarrow & \mathbb{Z}/2\,. \end{array}$$

The possibility of defining the Brouwer degree with values[1] in $GW(k)$ and the above cartesian square emphasizes one of our constant intuition in this paper and should be kept in mind: from the degree point of view, the (top horizontal) rank homomorphism corresponds to "taking care of the topology of the complex points" and the projection $GW(k) \to W(k)$ corresponds to "taking care of the topology of the real points". Indeed, given a real embedding $k \to \mathbb{R}$, with associated signature $W(k) \to \mathbb{Z}$, the signature of the degree of f is the degree of the associated map $f(\mathbb{R})\colon \mathbb{P}^1(\mathbb{R}) \to \mathbb{P}^1(\mathbb{R})$. This idea of taking care of these two topological intuitions at the same time is essential in the present work.

We do not pretend to be exhaustive in such a short paper; we have mostly emphasized the progress in unstable $\mathbb{A}^1$-homotopy theory and we will almost not address stable $\mathbb{A}^1$-homotopy theory.

Notations. We fix a base field k of any characteristic; Sm_k will denote the category of smooth quasi-projective k-schemes. Given a presheaf of sets on Sm_k, that is to say a functor $F\colon (\mathsf{Sm}_k)^{\mathrm{op}} \to \mathsf{Sets}$, and an essentially smooth k-algebra A, which means that A is the filtering union of its sub-k-algebras A_α which are smooth and finite type over k, we set $F(A) := \mathrm{colimit}_\alpha F(\mathrm{Spec}(A_\alpha))$. For instance, for each point $x \in X \in \mathsf{Sm}_k$ the local ring $\mathcal{O}_{X,x}$ of X at x as well as its henselization $\mathcal{O}^h_{X,x}$ are essentially smooth k-algebras.

Some history and acknowledgements. This work has its origin in my discussions and collaboration with V. Voevodsky [38]; I thank him very much for these discussions.

I thank J. Lannes for his influence and interest on my first proof of the Milnor conjecture on quadratic forms in [29], relying on Voevodsky's results and on the use of the Adams spectral sequence based on mod. 2 motivic cohomology. Since then I considerably simplified the topological argument in [33].

[1] Barge and Lannes have defined and studied a related degree from the set of naive $\mathbb{A}^1$-homotopy classes of k-morphisms $\mathbb{P}^1 \to \mathbb{P}^1$ to $GW(k)$, unpublished.

I also want to warmly thank M. Hopkins and M. Levine for their constant interest in this work -as well as related works-, for discussions and comments which helped me very much to simplify and improve some parts, and also for some nice collaborations on and around this subject during the past years.

Finally, I want to thank very much the Mathematics Institute as well as my colleagues of the University of Munich for their welcome.

2. A quick recollection on $\mathbb{A}^1$-homotopy

A convenient category of spaces. We will always consider that Sm_k is endowed with the Nisnevich topology [40], [38]. We simply recall the following characterization for a presheaf of sets on Sm_k to be a sheaf in this topology.

Proposition 2.1 ([38]). *A functor $F\colon (\mathsf{Sm}_k)^{\mathrm{op}} \to \mathsf{Sets}$ is a sheaf in the Nisnevich topology if and only if for any cartesian square in Sm_k of the form*

$$\begin{array}{ccc} W & \subset & V \\ \downarrow & & \downarrow \\ U & \subset & X \end{array} \tag{2.1}$$

where U is an open subscheme in X, the morphism $f\colon V \to X$ is étale and the induced morphism $(f^{-1}(X-U))_{\mathrm{red}} \to (X-U)_{\mathrm{red}}$ is an isomorphism, the map

$$F(X) \to F(U) \times_{F(W)} F(V)$$

is a bijection.

Squares like (2.1) are call *distinguished squares*. We denote by $\Delta^{\mathrm{op}}\mathsf{Shv}_k$ the category of simplicial sheaves of sets over Sm_k (in the Nisnevich topology); these objects will be just called "spaces" (this is slightly different from [54] where "space" only means a sheaf of sets, with no simplicial structure). This category contains the category Sm_k as the full subcategory of representable sheaves.

$\mathbb{A}^1$-weak equivalence and $\mathbb{A}^1$-homotopy category. Recall that a *simplicial weak equivalence* is a morphism of spaces $f\colon \mathcal{X} \to \mathcal{Y}$ such that each of its stalks

$$\mathcal{X}(\mathcal{O}^h_{X,x}) \to \mathcal{Y}(\mathcal{O}^h_{X,x})$$

at $x \in X \in \mathsf{Sm}_k$ is a weak equivalence of simplicial sets. Inverting these morphisms in $\Delta^{\mathrm{op}}\mathsf{Shv}_k$ yields the classical *simplicial homotopy category of sheaves* [10], [21]. The notion of $\mathbb{A}^1$-*weak equivalence* is generated in some natural way by that of simplicial weak equivalences and the projections $\mathcal{X} \times \mathbb{A}^1 \to \mathcal{X}$ for any space $\mathcal{X}$. Inverting the class of $\mathbb{A}^1$-weak equivalence yields now the $\mathbb{A}^1$-homotopy category $\mathsf{H}(k)$ [54], [38]. We denote by $\mathsf{H}_\bullet(k)$ the $\mathbb{A}^1$-homotopy category of pointed spaces.

The smash product with the simplicial circle S^1 induces the simplicial suspension functor $\Sigma\colon \mathsf{H}_\bullet(k) \to \mathsf{H}_\bullet(k)$, $\mathcal{X} \mapsto \Sigma(\mathcal{X})$. For a pointed morphism $f\colon \mathcal{X} \to \mathcal{Y}$ we may define the $\mathbb{A}^1$-homotopy fiber $\Gamma(f)$ together with an $\mathbb{A}^1$-fibration sequence $\Gamma(f) \to \mathcal{X} \to \mathcal{Y}$, which moreover induces for any other pointed space $\mathcal{Z}$ a long homotopy exact sequence (of pointed sets, groups, abelian groups as usual)

$$\begin{aligned}\cdots \to \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\Sigma(\mathcal{Z}), \mathcal{X}) \to \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\Sigma(\mathcal{Z}), \mathcal{Y}) \to \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\mathcal{Z}, \Gamma(f))\\ \to \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\mathcal{Z}, \mathcal{X}) \to \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\mathcal{Z}, \mathcal{Y}).\end{aligned}$$

In a dual way a distinguished square like (2.1) above is $\mathbb{A}^1$-homotopy cocartesian and induces corresponding Mayer–Vietoris type long exact sequences by mapping its vertices to $\mathcal{Z}$.

The geometric ideas on the Brouwer degree recalled in the introduction lead in general (for $d > n$) to the interpretation, due to Pontryagin, of the stable homotopy groups of spheres in terms of *parallelized cobordism groups*, and even more generally to the *Thom–Pontryagin construction* used by Thom to compute most of the cobordism rings. Recall that given a closed embedding $i\colon Z \hookrightarrow X$ between differentiable manifolds (with Z compact for simplicity) and a *tubular neighborhood* $Z \subset U \subset X$ of Z in X there is a pointed continuous map (indeed homeomorphism)

$$X/(X - U) \to Th(\nu_i) \tag{2.2}$$

to the Thom space (the one point compactification of the total space $E(\nu_i) \cong U$ of the normal bundle ν_i) which is independent up to pointed homotopy of the choices of the tubular neighborhood.

The choice of the topology on Sm_k (see [38]) was very much inspired by the this Thom–Pontryagin construction and the definition of the $\mathbb{A}^1$-homotopy category of smooth schemes over a base in [54], [38] allows to construct, for any closed immersion $i\colon Z \to X$ between smooth k-schemes, a pointed $\mathbb{A}^1$-weak equivalence $X/(X - Z) \to Th(\nu_i)$ [38], although no tubular neighborhood is available in general in algebraic geometry. In that case we get an $\mathbb{A}^1$-cofibration sequence

$$(X - Z) \to X \to Th(\nu_i).$$

Let $\mathcal{X}$ be a space. We let $\pi_0^{\mathbb{A}^1}(\mathcal{X})$ denote the associated sheaf (of sets) on Sm_k to the presheaf $U \mapsto \mathrm{Hom}_{\mathsf{H}(k)}(U, \mathcal{X})$. If moreover $\mathcal{X}$ is pointed, and $n \geq 1$ we denote by $\pi_n^{\mathbb{A}^1}(\mathcal{X})$ the sheaf on Sm_k associated to the presheaf $U \mapsto \mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\Sigma^n(U_+), \mathcal{X})$ (where U_+ means U together with a base point added outside), a sheaf of groups for $n = 1$, of abelian groups for $n \geq 2$.

It is also very useful for the intuition to recall from [38] the existence of the topological realization functors. When $\rho\colon k \to \mathbb{C}$ (resp. $k \to \mathbb{R}$) is a complex (resp. real) embedding there is a canonical functor $\mathsf{H}(k) \to \mathsf{H}$ to the usual homotopy category of C.W.-complexes, induced by sending $X \in \mathsf{Sm}_k$ to the set of complex points $X(\mathbb{C})$ (resp. real points $X(\mathbb{R})$) with its classical topology.

3. $\mathbb{A}^1$-homotopy and $\mathbb{A}^1$-homology: the basic theorems

We recall that everywhere in this paper the topology to be understood is the Nisnevich topology.

Strictly $\mathbb{A}^1$-invariant sheaves

Definition 3.1. 1) A presheaf of sets M on Sm_k is said to be $\mathbb{A}^1$-*invariant* if for any $X \in \mathsf{Sm}_k$, the map $M(X) \to M(X \times \mathbb{A}^1)$ induced by the projection $X \times \mathbb{A}^1 \to X$, is a bijection.

2) A sheaf of groups M is said to be *strongly* $\mathbb{A}^1$-*invariant* if for any $X \in \mathsf{Sm}_k$ and any $i \in \{0, 1\}$ the map $H^i(X; M) \to H^i(X \times \mathbb{A}^1; M)$ induced by the projection $X \times \mathbb{A}^1$, is a bijection.

3) A sheaf of abelian groups M is said to be *strictly* $\mathbb{A}^1$-*invariant* if for any $X \in \mathsf{Sm}_k$ and any $i \in \mathbb{N}$ the map $H^i(X; M) \to H^i(X \times \mathbb{A}^1; M)$ induced by the projection $X \times \mathbb{A}^1$ is a bijection.

These notions, except 2), appear in Voevodsky's study of cohomological properties of presheaves with transfers [55] and were extensively studied in [34] over a general base, though very few is known except when the base is a field. Hopefully, given a sheaf of abelian groups the *a priori* different properties 2) and 3) coincide.

Theorem 3.2 ([36]). *A sheaf of abelian groups which is strongly* $\mathbb{A}^1$-*invariant is strictly* $\mathbb{A}^1$-*invariant.*

This result can be used to simplify some of the proofs of [55]. Let us denote by Ab_k the abelian category of sheaves of abelian groups on Sm_k. Another easy application is that the full sub-category $\mathsf{Ab}_k^{\mathbb{A}^1} \subset \mathsf{Ab}_k$, consisting of strictly $\mathbb{A}^1$-invariant sheaves, is an abelian category for which the inclusion functor is exact. From Theorem 3.3 below, these strictly $\mathbb{A}^1$-invariant sheaves and their cohomology play in $\mathbb{A}^1$-algebraic topology the role played in classical algebraic topology by the abelian groups and the singular cohomology with coefficients in those.

The constant sheaf $\mathbb{Z}$, the sheaf represented by an abelian variety over k are examples of strictly $\mathbb{A}^1$-invariant sheaves, in fact the higher cohomology groups, $H^i_{Nis}(X; -)$, $i > 0$, for these sheaves automatically vanish. Another well known example is the multiplicative group $\mathbb{G}_m = \mathbb{A}^1 - \{0\}$. More elaborated examples were produced by Voevodsky over a perfect field: for each $\mathbb{A}^1$-homotopy invariant presheaf with transfers F its associated sheaf F_{Nis} a strictly $\mathbb{A}^1$-invariant sheaf [55]. In particular if F itself is an $\mathbb{A}^1$-homotopy invariant sheaf with transfers, it is strictly $\mathbb{A}^1$-invariant. By [12] these sheaves are very closely related to Rost's cycle modules [46], which also produce strictly $\mathbb{A}^1$-invariant sheaves, like the unramified Milnor K-theory sheaves introduced in [19]. There are other types of strictly $\mathbb{A}^1$-invariant sheaves given for instance by the unramified Witt groups $\underline{\mathbf{W}}$ as constructed in [42], or [36], as well as their subsheaves of unramified power of the fundamental ideal $\underline{\mathbf{I}}^n$ used in [33].

$\mathbb{A}^1$-homotopy sheaves

Theorem 3.3 ([36]). *Let $\mathcal{X}$ be a pointed space. Then the sheaf $\pi_1^{\mathbb{A}^1}(\mathcal{X})$ is strongly $\mathbb{A}^1$-invariant, and the sheaves $\pi_n^{\mathbb{A}^1}$, for $n \geq 2$, are strictly $\mathbb{A}^1$-invariant.*

Curiously enough, we are unable to prove that the sheaf $\pi_0^{\mathbb{A}^1}(\mathcal{X})$ is $\mathbb{A}^1$-invariant, though it is true in all the cases we can compute.

Remark 3.4. One of the main tool used in the proof of the Theorem 3.3 is the presentation Lemma of Gabber [14] as formalized in [11]. Then a "non-abelian" variant of [11] and ideas from [46] lead to the result. In fact one can give a quite concrete description of a sheaf of groups which is strongly $\mathbb{A}^1$-invariant [36].

A pointed space $\mathcal{X}$ such that the sheaves $\pi_i^{\mathbb{A}^1}(\mathcal{X})$ vanish for $i \leq n$ will be called n-$\mathbb{A}^1$-connected. In case $n = 0$ we simply say $\mathbb{A}^1$-connected.

Corollary 3.5 (Unstable $\mathbb{A}^1$-connectivity theorem). *Let $\mathcal{X}$ be a pointed space and n be an integer ≥ 0 such that $\mathcal{X}$ is* simplicially *n-connected. Then it is n-$\mathbb{A}^1$-connected.*

This result was only known in the case $n = 0$ in [38], over a general base. As a consequence, the simplicial suspension of an $(n-1)$-$\mathbb{A}^1$-connected pointed space is n-$\mathbb{A}^1$-connected.

The main example of a simplicially n-connected space is the $(n+1)$-th simplicial suspension of a pointed space.

For n and i two natural numbers we set $S^n(i) = (S^1)^{\wedge(n)} \wedge (\mathbb{G}_m)^{\wedge i}$ where $\wedge$ denotes the smash-product. Observe that these are actually mapped to spheres (up to homotopy) through any topological realization functors (real or complex). Note also the following isomorphisms in $\mathsf{H}_\bullet(k)$: $\mathbb{A}^n - \{0\} \cong S^{(n-1)}(n)$ and $(\mathbb{P}^1)^{\wedge n} \cong S^1 \wedge (\mathbb{A}^n - \{0\}) \cong S^n(n)$.

From the previous Corollary $S^n(i)$ is $(n-1)$-$\mathbb{A}^1$-connected. Actually we will see below that it is exactly $(n-1)$-$\mathbb{A}^1$-connected, as $\pi_n^{\mathbb{A}^1}(S^n(i))$ is always non trivial. The $\mathbb{A}^1$-connectivity corresponds to the connectivity of the space of real points.

$\mathbb{A}^1$-fundamental group and universal $\mathbb{A}^1$-covering.

An $\mathbb{A}^1$-trivial cofibration $\mathcal{A} \to \mathcal{B}$ is a monomorphism between spaces which is also an $\mathbb{A}^1$-weak equivalence. The following definition is the obvious analogue of the definition of a covering in topology:

Definition 3.6. An $\mathbb{A}^1$-covering $\mathcal{Y} \to \mathcal{X}$ is a morphism of spaces which has the unique right lifting property with respect to $\mathbb{A}^1$-trivial cofibrations. This means that given any commutative square of spaces

$$\begin{array}{ccc} \mathcal{A} & \longrightarrow & \mathcal{Y} \\ \downarrow & & \downarrow \\ \mathcal{B} & \longrightarrow & \mathcal{X} \end{array}$$

in which $\mathcal{A} \to \mathcal{B}$ is an $\mathbb{A}^1$-trivial cofibration, there exists one and exactly one morphism $\mathcal{B} \to \mathcal{Y}$ which makes the whole diagram commutative.

Example 3.7. 1) Any finite étale covering $Y \to X$ between smooth k-varieties, in characteristic 0, is an $\mathbb{A}^1$-covering. Any Galois étale covering $Y \to X$ with Galois group of order prime to the characteristic of k is an $\mathbb{A}^1$-covering.

2) Any $\mathbb{G}_m$-torsor $\mathcal{Y} \to \mathcal{X}$ is an $\mathbb{A}^1$-covering. Remember to think about the real points! A $\mathbb{G}_m$-torsor gives (up to homotopy) a $\mathbb{Z}/2$-covering.

Theorem 3.8. *Any pointed $\mathbb{A}^1$-connected space $\mathcal{X}$ admits a universal pointed $\mathbb{A}^1$-covering $\tilde{\mathcal{X}} \to \mathcal{X}$ in the category of pointed coverings of $\mathcal{X}$. The fiber of this universal $\mathbb{A}^1$-covering at the base point is isomorphic to $\pi_1^{\mathbb{A}^1}(\mathcal{X})$ and $\tilde{\mathcal{X}} \to \mathcal{X}$ is (up to canonical isomorphism) the unique pointed $\mathbb{A}^1$-covering with $\tilde{\mathcal{X}}$ being 1-$\mathbb{A}^1$-connected.*

Remark 3.9. A pointed $\mathbb{A}^1$-connected smooth k-scheme (X, x) admits no non-trivial étale pointed covering. Thus the $\pi_1^{\mathbb{A}^1}$ is in some sense orthogonal to the étale one and gives a more combinatorial information, as shown by the example of the $\mathbb{P}^n$'s below. On the other hand the pointed étale coverings always come from the $\pi_0^{\mathbb{A}^1}$: for instance an abelian variety X is discreet, in the sense that $\pi_0^{\mathbb{A}^1}(X) = X$, and have huge étale π_1. We did not try to further study the $\mathbb{A}^1$-fundamental groupoid which cares about both aspects, the combinatorial and the étale.

Lemma 3.10. *Let $n \geq 2$. The canonical $\mathbb{G}_m$-torsor*

$$(\mathbb{A}^{n+1} - \{0\}) \to \mathbb{P}^n$$

is the universal covering of $\mathbb{P}^n$. As a consequence the morphism $\pi_1^{\mathbb{A}^1}(\mathbb{P}^n) \to \mathbb{G}_m$ is an isomorphism.

Indeed, $\mathbb{A}^{n+1} - \{0\}$ is 1-$\mathbb{A}^1$-connected. For $n = 1$ the problem is that $\mathbb{A}^2 - \{0\}$ is no longer 1-$\mathbb{A}^1$-connected. See the next section for more information.

$\mathbb{A}^1$-derived category, $\mathbb{A}^1$-homology and Hurewicz Theorem. Let us denote by $\mathbb{Z}(\mathcal{X})$ the free abelian sheaf generated by a space $\mathcal{X}$ and by $C_*(\mathcal{X})$ its the associated chain complex; if moreover $\mathcal{X}$ is pointed, let us denote by $\mathbb{Z}_\bullet(\mathcal{X}) = \mathbb{Z}(\mathcal{X})/\mathbb{Z}$ and $\tilde{C}_*(\mathcal{X}) = C_*(\mathcal{X})/\mathbb{Z}$ the reduced versions obtained by collapsing the base point to 0.

We may perform in the derived category of chain complexes in Ab_k exactly the same process as for spaces and define the class of $\mathbb{A}^1$-weak equivalences, rather $\mathbb{A}^1$-quasi isomorphisms; these are generated by quasi-isomorphisms and collapsing $\mathbb{Z}_\bullet(\mathbb{A}^1)$ to 0. Formally inverting these morphisms yields the *$\mathbb{A}^1$-derived category* $\mathsf{D}_{\mathbb{A}^1}(k)$ of k [34]. The functor $\mathcal{X} \mapsto C_*(\mathcal{X})$ obviously induces a functor $\mathsf{H}(k) \to \mathsf{D}_{\mathbb{A}^1}(k)$ which admits a right adjoint given by the usual Eilenberg–MacLane functor $K : \mathsf{D}_{\mathbb{A}^1}(k) \to \mathsf{H}(k)$.

As for spaces, one may define $\mathbb{A}^1$-homology sheaves of a chain complex C_*. An abelian version of Theorem 3.3 implies that for any complex C_* these $\mathbb{A}^1$-homology sheaves are strictly $\mathbb{A}^1$-invariant [36], [34].

Definition 3.11. For a space $\mathcal{X}$ and for each integer $n \in \mathbb{Z}$, we let $\mathbb{H}_n^{\mathbb{A}^1}(X)$ denote the n-th $\mathbb{A}^1$-homology sheaf of $C_*(\mathcal{X})$ and call $\mathbb{H}_*^{\mathbb{A}^1}(\mathcal{X})$ the $\mathbb{A}^1$-homology of $\mathcal{X}$ (with integral coefficients). In case $\mathcal{X}$ is pointed, we let $\tilde{\mathbb{H}}_*^{\mathbb{A}^1}(\mathcal{X})$ denote the reduced version obtained by collapsing the base point to 0.

Observe that these $\mathbb{A}^1$-homology sheaves are strictly $\mathbb{A}^1$-invariant and that $\mathbb{H}_i^{\mathbb{A}^1}(\mathcal{X}) = 0$ for $i < 0$ by the abelian analogue of Corollary 3.5. As a consequence for a space $\mathcal{X}$ the sheaf $\mathbb{H}_0^{\mathbb{A}^1}(\mathcal{X})$ is the *free strictly* $\mathbb{A}^1$*-invariant sheaf* generated by $\mathcal{X}$. These sheaves play a fundamental role in $\mathbb{A}^1$-algebraic topology. For instance we have suspension isomorphisms $\tilde{\mathbb{H}}_*^{\mathbb{A}^1}(S^n(i)) \cong \tilde{\mathbb{H}}_{*-n}^{\mathbb{A}^1}((\mathbb{G}_m)^{\wedge i})$ for our spheres $S^n(i)$. In particular the first *a priori* non trivial sheaf is $\tilde{\mathbb{H}}_n^{\mathbb{A}^1}(S^n(i)) \cong \tilde{\mathbb{H}}_0^{\mathbb{A}^1}((\mathbb{G}_m)^{\wedge i})$. We will compute these sheaves in the next section in terms of Milnor–Witt K-theory.

The computation of the higher $\mathbb{A}^1$-homology sheaves is at the moment highly non trivial and mysterious[2].

Remark 3.12. There exists a natural morphism of sheaves $\mathbb{H}_n^{\mathbb{A}^1}(X;\mathbb{Z}) \to \mathbb{H}_n^S(X)$ where the right hand side denotes Suslin–Voevodsky singular homology sheaves [52], [55]. In general, this is not an isomorphism. More generally let $\mathsf{DM}(k)$ be Voevodsky's triangulated category of motives [56]. Then there exists a canonical functor of "adding transfers"

$$\mathsf{D}_{\mathbb{A}^1}(k) \to \mathsf{DM}(k).$$

It is not an equivalence. One explanation is given by the (pointed) algebraic Hopf map:

$$\eta\colon \mathbb{A}^2 - \{0\} \to \mathbb{P}^1.$$

The associated morphism on $\mathbb{H}_1^{\mathbb{A}^1}$ defines a morphism[3]:

$$\eta\colon \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \otimes_{\mathbb{A}^1} \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \cong \mathbb{H}_1^{\mathbb{A}^1}(\mathbb{A}^2 - \{0\}) \to \mathbb{H}_1^{\mathbb{A}^1}(\mathbb{P}^1) \cong \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m).$$

The latter is never nilpotent (use the same argument as in the proof of Theorem 4.7). On the other hand, the computation of the motive of $\mathbb{P}^2$, which is the cone of η, shows that $\mathbb{P}^1 \to \mathbb{P}^2$ admits a retraction in $\mathsf{DM}(k)$ and thus that the image of η in $\mathsf{DM}(k)$ is the zero morphism.

Theorem 3.13 (Hurewicz Theorem, [36]). *Let $\mathcal{X}$ be a pointed $\mathbb{A}^1$-connected space. Then the Hurewicz morphism*

$$\pi_1^{\mathbb{A}^1}(\mathcal{X}) \to \mathbb{H}_1^{\mathbb{A}^1}(\mathcal{X})$$

[2] We do not know any example which does not use the Bloch–Kato conjecture.

[3] Here for sheaves M and N, we denote by $M \otimes_{\mathbb{A}^1} N$ the $\mathbb{H}_0^{\mathbb{A}^1}$ of the sheaf $M \otimes N$, and call it the $\mathbb{A}^1$-tensor product.

is the universal morphism from $\pi_1^{\mathbb{A}^1}(\mathcal{X})$ *to a strictly* $\mathbb{A}^1$*-invariant sheaf*[4]*. If moreover* $\mathcal{X}$ *is* $(n-1)$*-connected for some* $n \geq 2$ *then the Hurewicz morphism*

$$\pi_i^{\mathbb{A}^1}(\mathcal{X}) \to \mathbb{H}_i^{\mathbb{A}^1}(\mathcal{X})$$

is an isomorphism for $i \leq n$ *and an epimorphism for* $i = (n+1)$.

We now may partly realize our program of proving the analogue of Theorem 1.1. Given a sphere $S^n(i)$ with $n \geq 2$, we have $\pi_m^{\mathbb{A}^1}(S^n(i)) = 0$ for $m < n$ and

$$\pi_n^{\mathbb{A}^1}(S^n(i)) \cong \tilde{\mathbb{H}}_0^{\mathbb{A}^1}((\mathbb{G}_m)^{\wedge n}) \cong \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m)^{\otimes_{\mathbb{A}^1}(n)}.$$

In the next section we will describe those sheaves.

Remark 3.14. Of course, the Hurewicz Theorem has a lot of classical consequences. We do not mention them here, see [36].

4. $\mathbb{A}^1$-homotopy and $\mathbb{A}^1$-homology: computations involving Milnor–Witt K-theory

Milnor–Witt K-theory of fields. The following definition was obtained in collaboration with Mike Hopkins.

Definition 4.1. Let F be a commutative field. The Milnor–Witt K-theory $K_*^{MW}(F)$ of F is the graded associative ring generated by the symbols $[u]$, for each unit $u \in F^\times$, of degree $+1$, and η of degree -1 subject to the following relations:

(1) (Steinberg relation) For each $a \in F^\times - \{1\}$, one has $[a].[1-a] = 0$.

(2) For each pair $(a,b) \in (F^\times)^2$ one has $[ab] = [a] + [b] + \eta.[a].[b]$.

(3) For each $a \in F^\times$, one has $[a].\eta = \eta.[a]$.

(4) One has $\eta^2.[-1] + 2\eta = 0$.

This Milnor–Witt K-theory groups were introduced by the author in a different complicated way. The previous one, is very simple and natural (but maybe the 4-th relation which will be explained below): all the relations easily come from natural $\mathbb{A}^1$-homotopies, see Theorem 4.8.

The quotient $K_*^{MW}(F)/\eta$ of the Milnor–Witt K-theory of F by η is the *Milnor K-theory* $K_*^M(F)$ of F as defined in [26]; indeed after η is killed, the symbol $[a]$ becomes additive and there is only the Steinberg relation.

For any unit $a \in F^\times$, set $\langle a \rangle = \eta[a] + 1 \in K_0^{MW}(F)$. One can show that $[1] = 0$, $\langle 1 \rangle = 1$ and $\langle ab \rangle = \langle a \rangle \langle b \rangle$. Set $\varepsilon := -\langle -1 \rangle$ and $h = 1 + \langle -1 \rangle$. Observe that $h = \eta.[-1] + 2$ and the fourth relation can be written $\eta\varepsilon = \eta$ or equivalently $\eta \,.\, h = 0$.

[4] it is not yet known whether this is the abelianization nor an epimorphism

This η will be interpreted below in term of the algebraic Hopf map (see also Remark 3.12 above). Observe that the relation $\eta^2.[-1] + 2\eta = 0$ is compatible with the complex points (where $[-1] = 0$ and stably $2.\eta = 0$) and the real points (where $[-1] = -1$, $\eta = 2$ and $-2^2 + 2 \times 2 = 0$).

It is natural to call the quotient ring $K^{MW}_*(F)/h$ the *Witt K-theory* of F and to denote it by $K^W_*(F)$. The mod 2-Milnor K-theory $k_*(F) := K^M_*(F)/2$ is thus also the mod η Witt K-theory $K^W_*(F)/\eta = K^{MW}_*(F)/(h, \eta)$.

It is not hard to check that $K^{MW}_0(F)$ admits the following presentation as an abelian group: a generator $\langle u\rangle$ for each unit of $F^\times$ and the relations of the form: $\langle u(v^2)\rangle = \langle u\rangle$, $\langle u\rangle + \langle v\rangle = \langle u + v\rangle + \langle (u + v)uv\rangle$ if $(u + v) \neq 0$ and $\langle u\rangle + \langle -u\rangle = 1 + \langle -1\rangle$. Moreover one checks that the morphism $\eta^n : K^{MW}_0(F) \to K^{MW}_{-n}(F)$ induces an isomorphism $K^W_0(F) \cong K^{MW}_{-n}(F)$ for $n > 0$. Thus in particular $K^{MW}_*(F)[\eta^{-1}] \to K^W_0(F)[\eta, \eta^{-1}]$ is an isomorphism.

Remark 4.2. In the above presentation of $K^{MW}_0(F)$ one recognizes the presentation of the Grothendieck–Witt ring $GW(F)$, see [47] in the case of characteristic $\neq 2$ and [27] in the general case. The element h becomes the hyperbolic plane. The quotient group (actually a ring) $K^W_0(F) = GW(F)/h$ is exactly the Witt ring $W(F)$ of F.

Let us define the *fundamental ideal* $I(F)$ of $K^W_0(F)$ to be the kernel of the mod 2 rank homomorphism $K^W_0(F) \to \mathbb{Z}/2$. Set $I^*(F) = \bigoplus_{n\in\mathbb{Z}} I^n(F)$ (with the convention $I^n(F) = K^W_0(F)$ for $n \leq 0$). We observe that the obvious correspondence $[u] \mapsto \langle u\rangle - 1 \in I(F)$ induces an (epi)morphism

$$S_F : K^W_*(F) \to I^*(F)$$

where η acts through the inclusions $I^n(F) \subset I^{n-1}(F)$. Killing η in this morphism yields the Milnor morphism [26]:

$$s_F : k_*(F) \to i^*(F) \tag{4.1}$$

where $i^*(F)$ denotes $\oplus I^n(F)/I^{(n+1)}(F)$.

Theorem 4.3 ([32]). *For any field F of characteristic $\neq 2$ the homomorphism*

$$S_F : K^W_*(F) \to I^*(F)$$

is an isomorphism.

This statement cannot be trivial as it implies the Milnor conjecture on quadratic forms that morphism (4.1) is an isomorphism. This statement is a reformulation of [1] and thus uses the proof of the Milnor conjecture on mod 2 Galois cohomology by Voevodsky [57], [41].

As a consequence we obtained in [32] that the commutative square of graded rings,

$$\begin{array}{ccc} K^{MW}_*(F) & \longrightarrow & K^M_*(F) \\ \downarrow & & \downarrow \\ K^W_*(F) & \longrightarrow & k_*(F) \end{array} \tag{4.2}$$

is cartesian (for a field of characteristic $\neq 2$).

Remark 4.4. 1) Using Kato's proof [20] of the analogue of the Milnor conjecture in characteristic 2, we can also show the previous result holds in characteristic 2.

2) The fiber products of the form $I^n(F) \times_{i^n(F)} K^M_n(F)$ where considered in [5] in characteristic not 2.

For $n \geq 1$ we simply set

$$\mathbb{Z}_{\mathbb{A}^1}(n) := \tilde{\mathbb{H}}^{\mathbb{A}^1}_0((\mathbb{G}_m)^{\wedge n})$$

for the free (reduced) strictly $\mathbb{A}^1$-invariant sheaf on $(\mathbb{G}_m)^{\wedge n}$. The Hopf morphism $\eta\colon \mathbb{A}^2 - \{0\} \to \mathbb{P}^1$ induces on $\tilde{\mathbb{H}}^{\mathbb{A}^1}_1$ a morphism of the form $\eta\colon \mathbb{Z}_{\mathbb{A}^1}(2) \to \mathbb{Z}_{\mathbb{A}^1}(1)$. Observe that $\tilde{\mathbb{H}}^{\mathbb{A}^1}_0((\mathbb{G}_m)^{\wedge 0}) = \mathbb{H}^{\mathbb{A}^1}_0(\mathrm{Spec}(k)) = \mathbb{Z}$ but that we did not set $\mathbb{Z}_{\mathbb{A}^1}(0) = \mathbb{Z}$. We will in fact extend this family of sheaves $\mathbb{Z}_{\mathbb{A}^1}(n)_{n\geq 1}$ to integers $n \leq 0$ using a construction of Voevodsky.

Given a presheaf of pointed sets M one defines the *pointed $\mathbb{G}_m$-loop space* M_{-1} on M so that for $X \in \mathsf{Sm}_k$, $M_{-1}(X)$ is the "Kernel" of the restriction through the unit section $M(X \times \mathbb{G}_m) \to M(X)$. If M is a sheaf of abelian groups, so is M_{-1}. We may iterate this construction to get M_n for $n < 0$; we set, for $n \leq 0$

$$\mathbb{Z}_{\mathbb{A}^1}(n) = \mathbb{Z}_{\mathbb{A}^1}(1)_{n-1}.$$

The canonical morphism $\mathbb{Z} \to \mathbb{Z}_{\mathbb{A}^1}(0)$ is far from being an isomorphism. The tensor product (and internal Hom) defines natural pairings $\mathbb{Z}_{\mathbb{A}^1}(n) \otimes \mathbb{Z}_{\mathbb{A}^1}(m) \to \mathbb{Z}_{\mathbb{A}^1}(n+m)$ for any integers $(n, m) \in \mathbb{Z}^2$. The element η becomes now an element $\eta \in \mathbb{Z}_{\mathbb{A}^1}(-1)(k)$. Any unit $u \in F^\times$ in a separable field extension $F|k$, viewed as an element in $\mathbb{G}_m(F)$ defines an element $[u] \in \mathbb{Z}_{\mathbb{A}^1}(1)(F)$.

The following result own very much to the definition of the Milnor–Witt K-theory found with Hopkins:

Theorem 4.5 ([28]). *For any separable field extension $F|k$, the symbols $[u] \in \mathbb{Z}_{\mathbb{A}^1}(1)(F)$, for any $u \in F^\times$, and $\eta \in \mathbb{Z}_{\mathbb{A}^1}(-1)(F)$, satisfy the* 4 *relations of Definition* 4.1 *in the graded ring $\mathbb{Z}_{\mathbb{A}^1}(*)(F)$. We thus obtain a canonical homomorphism of graded rings*

$$\Theta_*(F)\colon K^{MW}_*(F) \to \mathbb{Z}_{\mathbb{A}^1}(*)(F).$$

The Steinberg relation (1) is a consequence of the following nice result of P. Hu and I. Kriz [17]. Consider the canonical closed immersion $\mathbb{A}^1 - \{0, 1\} \hookrightarrow \mathbb{G}_m \times \mathbb{G}_m$, $x \mapsto (x, 1-x)$. Then its (unreduced) suspension $\tilde{\Sigma}^1(\mathbb{A}^1 - \{0, 1\}) \to \Sigma^1(\mathbb{G}_m \times \mathbb{G}_m)$ composed with $\Sigma^1(\mathbb{G}_m \times \mathbb{G}_m) \to \Sigma^1(\mathbb{G}_m \wedge \mathbb{G}_m)$ is trivial in $\mathsf{H}_\bullet(k)$. Applying $\mathbb{H}_1^{\mathbb{A}^1}$ yields the Steinberg relation.

The last 3 relations are consequences of the following fact: let $\mu\colon \mathbb{G}_m \times \mathbb{G}_m \to \mathbb{G}_m$ denote the product morphism of the group scheme $\mathbb{G}_m$, then the induced morphism on $\mathbb{H}_1^{\mathbb{A}^1}$, $\mathbb{Z}_{\mathbb{A}^1}(1) \oplus \mathbb{Z}_{\mathbb{A}^1}(1) \oplus \mathbb{Z}_{\mathbb{A}^1}(2) \to \mathbb{Z}_{\mathbb{A}^1}(1)$ is of the form $\mathrm{Id}_{\mathbb{Z}_{\mathbb{A}^1}(1)} \oplus \mathrm{Id}_{\mathbb{Z}_{\mathbb{A}^1}(1)} \oplus \eta$. The relation (2) follows clearly form this fact. The relations (3) and (4) follow from the commutativity of μ. □

Unramified Milnor–Witt K-theory and the main computation. We next define for each $n \in \mathbb{Z}$ an explicit sheaf $\underline{\mathbf{K}}_n^{MW}$ called the sheaf of *unramified Milnor–Witt K-theory* in weight n. To do this, let us give some recollection. For the Milnor K-theory [26], for any discrete valuation v on a field F, with valuation ring $\mathcal{O}_v \subset F$, residue field $\kappa(v)$, one can define a unique homomorphism (of graded groups)

$$\partial_v\colon K_*^M(F) \to K_{*-1}^M(\kappa(v))$$

called "residue" homomorphism, such that

$$\partial_v(\{\pi\}\{u_2\}\dots\{u_n\}) = \{\overline{u_2}\}\dots\{\overline{u_n}\}$$

for any uniformizing element π (of v) and units $u_i \in \mathcal{O}_v^\times$, and where $\bar{u}$ denotes the image of $u \in \mathcal{O}_v \cap F^\times$ in $\kappa(v)$.

In the same way, given a uniformizing element π, one can define a residue morphism

$$\partial_v^\pi\colon K_*^{MW}(F) \to K_{*-1}^{MW}(\kappa(v))$$

satisfying the formula:

$$\partial_v^\pi([\pi].[u_2]\dots[u_n]) = [\overline{u_2}]\dots[\overline{u_n}].$$

However, one important feature is that in the case of Milnor K-theory, these residues do not depend on the choice of π, only on the valuation, but in the case of Milnor–Witt K-theory, they do depend on the choice of π: for $u \in \mathcal{O}^\times$, as one has $\partial_v^\pi([u.\pi]) = \partial_v^\pi([\pi]) + \eta.[\bar{u}] = 1 + \eta.[\bar{u}]$.

To make this residue homomorphism "canonical" (see [5], [6], [48] for instance), one defines for a field κ and a one dimensional κ-vector space L, twisted Milnor–Witt K-theory groups: $K_*^{MW}(\kappa; L) = K_*^{MW}(\kappa) \otimes_{\mathbb{Z}[\kappa^\times]} \mathbb{Z}[L - \{0\}]$, where the group ring $\mathbb{Z}[\kappa^\times]$ acts through $u \mapsto \langle u\rangle$ on $K_*^{MW}(\kappa)$ and through multiplication on $\mathbb{Z}[L - \{0\}]$. The canonical residue homomorphism is of the following form

$$\partial_v\colon K_*^{MW}(F) \to K_{*-1}^{MW}(\kappa(v); m_v/(m_v)^2)$$

with $\partial_v([\pi].[u_2]\dots[u_n]) = [\overline{u_2}]\dots[\overline{u_n}]\otimes\overline{\pi}$, where $m_v/(m_v)^2$ is the cotangent space at v (a one dimensional $\kappa(v)$-vector space).

Using these residue homomorphisms, one may define for any smooth k-scheme $X \in \mathsf{Sm}_k$, irreducible say, with function field K, and any $n \in \mathbb{Z}$, the group $\underline{\mathbf{K}}^{MW}(X)$ of unramified Milnor–Witt K-theory in weight n as the kernel of the (locally finite) sum of the residues at points x of codimension 1, viewed as discrete valuations on K:

$$K_n^{MW}(K) \xrightarrow{\Sigma_x \partial_x} \bigoplus_{x \in X^{(1)}} K_{n-1}^{MW}(\kappa(x); m_x/(m_x)^2)$$

and extends this to a sheaf $X \mapsto \underline{\mathbf{K}}_n^{MW}(X)$.

Example 4.6. 1) In [18] Kato considered first the sheaves of unramified Milnor K-theory $\underline{\mathbf{K}}_n^M$ defined exactly in the same way on the Zariski site of X. It was turned into a strictly $\mathbb{A}^1$-invariant sheaf (on Sm_k) by Rost in [46].

2) One may also define unramified Witt K-theory $\underline{\mathbf{K}}_n^W$, unramified mod 2 Milnor K-theory $\underline{\mathbf{k}}_n$ in the same way, etc.

These types of cohomology theories easily give the non nilpotence of η:

Theorem 4.7. *Let $n \geq 1$ and $i \geq 1$ be natural numbers. The n-th suspension in $\mathsf{H}_\bullet(k)$*

$$\Sigma^n(\eta^i)\colon S^{n+1}(i+1) \to S^{n+1}(1)$$

of the i-th iteration of the Hopf map $\eta\colon S^1(2) \to S^1(1)$, is never trivial. Thus the algebraic Hopf map is not stably nilpotent.

This is trivial if one has a real embedding as $\eta(\mathbb{R})$ is the degree 2 map. In general, one uses the cohomology theory $H^*(-; \underline{\mathbf{K}}_*^{MW}[\eta^{-1}])$, in which η induces an isomorphism. To conclude remember that $K_*^{MW}(k)[\eta^{-1}] = K_0^W(k)[\eta, \eta^{-1}]$ and that $K_0^W(k)$ is never 0 (for k algebraically closed it is $\mathbb{Z}/2$).

We can now state our main computational result. Any strictly $\mathbb{A}^1$-invariant sheaf M has residue homomorphisms (see [34] for instance) and one proves that the homomorphism of Theorem 4.8

$$\Theta_*(F)\colon K_*^{MW}(F) \to \mathbb{Z}_{\mathbb{A}^1}(*)(F)$$

is compatible with residues. Thus (by [33, A.1] it induces a morphism of sheaves

$$\Theta_*\colon \underline{\mathbf{K}}_*^{MW} \to \mathbb{Z}_{\mathbb{A}^1}(*). \tag{4.3}$$

Theorem 4.8 ([28]). *The above morphism* (4.3) *is an isomorphism.*

We observe that the product $\mathbb{G}_m \wedge \underline{\mathbf{K}}_n^{MW} \to \underline{\mathbf{K}}_1^{MW} \wedge \underline{\mathbf{K}}_n^{MW} \to \underline{\mathbf{K}}_{n+1}^{MW}$ induces an isomorphism $\underline{\mathbf{K}}_n^{MW} \cong (\underline{\mathbf{K}}_{n+1}^{MW})_{-1}$. We deduce the existence for each $n > 0$, each $i > 0$, of a canonical $\mathsf{H}_\bullet(k)$-morphism

$$S^n(i) \to K(\underline{\mathbf{K}}_i^{MW}, n). \tag{4.4}$$

Some consequences and applications. The previous result and the Hurewicz Theorem imply:

Theorem 4.9. *For any $n \geq 2$, any $i > 0$:*

1) *The morphism* (4.4) *induces an isomorphism*

$$\pi_n^{\mathbb{A}^1}(S^n(i)) \cong \underline{\mathbf{K}}_i^{MW}.$$

2) *For any $m \in \mathbb{N}$, any $j \in \mathbb{N}$, the previous isomorphism induces canonical isomorphisms*

$$\mathrm{Hom}_{\mathsf{H}(k)}(S^m(j), S^n(i)) \cong \begin{cases} 0 & \text{if } m < n, \\ K_{i-j}^{MW}(k) & \text{if } m = n. \end{cases}$$

In case $i = 0, \pi_n^{\mathbb{A}^1}(S^n) = \mathbb{Z}$ and $\mathrm{Hom}_{\mathsf{H}(k)}(S^m(j), S^n) = \begin{cases} 0 \text{ if } m < n \quad \text{or } j \neq 0, \\ \mathbb{Z} \text{ if } m = n \quad \text{and } j = 0. \end{cases}$

In general, for $n = 1$ the question is much harder, and in fact unknown. We only know $\pi_1^{\mathbb{A}^1}(S^1(i))$ in the cases $i = 0, 1, 2$. For $i = 0$, $\pi_1^{\mathbb{A}^1}(S^1(i))(S^1) = \mathbb{Z}$.

For $i = 2$, as $\mathrm{SL}_2 \to \mathbb{A}^2 - \{0\} \cong S^1(2)$ is an $\mathbb{A}^1$-weak equivalence, the sphere $S^1(2)$ is an h-space and (by Hurewicz Theorem and Theorem 3.2) $\pi_1^{\mathbb{A}^1}(S^1(2)) = \mathbb{H}_1^{\mathbb{A}^1}(S^1(2)) = \underline{\mathbf{K}}_2^{MW}$. In fact the universal $\mathbb{A}^1$-covering given by Theorem 3.8 admits a group structure and we thus get an extension of sheaves of groups (in fact in the Zariski topology as well)

$$0 \to \underline{\mathbf{K}}_2^{MW} \to \widetilde{\mathrm{SL}}_2 \to \mathrm{SL}_2 \to 1.$$

This is a central extension which also arises in the following way. Let $B(\mathrm{SL}_2)$ denote the simplicial classifying space of SL_2. Then the canonical cohomology class $\Sigma(\mathrm{SL}_2) \cong S^2(2) \to K(\underline{\mathbf{K}}_2^{MW}, 2)$ can be uniquely extended to a $\mathsf{H}_\bullet(k)$-morphism:

$$B(\mathrm{SL}_2) \to K(\underline{\mathbf{K}}_2^{MW}, 2)$$

because the quotient $B(\mathrm{SL}_2)/\Sigma(\mathrm{SL}_2)$ is 3-$\mathbb{A}^1$-connected. It is well-known that such an element in $H^2(B(\mathrm{SL}_2); \underline{\mathbf{K}}_2^{MW})$ corresponds to a central extension of sheaves as above. It is the universal $\mathbb{A}^1$-covering for SL_2.

Remark 4.10. 1) In view of [13] it should be interesting to determine the possible $\pi_1^{\mathbb{A}^1}$ of linear algebraic groups.

2) A. Suslin has computed in [49] the group $H_2(\mathrm{SL}_2(k))$ for most field k and found exactly $\underline{\mathbf{K}}_2^{MW}(k) = I^2(k) \times_{i^2(k)} K_2^M(k)$. This computation has clearly influenced our work.

To understand $\pi_1^{\mathbb{A}^1}(\mathbb{P}^1)$ we use the $\mathbb{A}^1$-fibration sequence

$$\mathbb{A}^2 - \{0\} \to \mathbb{P}^1 \to \mathbb{P}^\infty \tag{4.5}$$

which, using the long exact sequence of $\mathbb{A}^1$-homotopy sheaves, gives a short exact sequence of the form:

$$1 \to \underline{\mathbf{K}}_2^{MW} \to \pi_1^{\mathbb{A}^1}(\mathbb{P}^1) \to \mathbb{G}_m \to 1$$

because $\underline{\mathbf{K}}_2^{MW} = \pi_1^{\mathbb{A}^1}(\mathbb{A}^2 - \{0\})$ and because $\mathbb{P}^\infty \cong B(\mathbb{G}_m)$ has only non-trivial $\pi_1^{\mathbb{A}^1}$ equal to $\mathbb{G}_m$. This extension of (sheaves of) groups can be completely explicited [36]. In particular $\pi_1^{\mathbb{A}^1}(\mathbb{P}^1)$ is non abelian!

The Brouwer degree. Now we can deduce as particular case of Theorem 4.9 what we announced in the introduction.

Corollary 4.11. *For any $n \geq 2$, any $i > 0$, the degree morphism induced by the morphism* (4.4)

$$\mathrm{Hom}_{\mathsf{H}(k)}(S^n(i), S^n(i)) \to K_0^{MW}(k)$$

is an isomorphism. As a consequence, the endomorphism ring of the $\mathbb{P}^1$-sphere spectrum $\mathbb{S}^0$, which by definition is

$$\pi_0^{\mathbb{A}^1}(\mathbb{S}^0) = \mathrm{colim}_{n\to\infty} \mathrm{Hom}_{\mathsf{H}(k)}(S^n(n), S^n(n)),$$

is isomorphic to the Grothendieck–Witt ring $GW(k) = K_0^{MW}(k)$ of k (see [31]*,* [30] *for the case of a perfect field of characteristic $\neq 2$).*

When $n = 1$, $i = 1$, $S^1(1) \cong \mathbb{P}^1$, using the $\mathbb{A}^1$-fibration sequence (4.5) one may entirely describe $\mathrm{Hom}_{\mathsf{H}(k)}(\mathbb{P}^1, \mathbb{P}^1)$ [36]. One may check the morphism $\mathbb{P}^1 \to K(\underline{\mathbf{K}}_1^{MW}, 1)$ induces a degree morphism $\mathrm{Hom}_{\mathsf{H}_\bullet(k)}(\mathbb{P}^1, \mathbb{P}^1) \to K_0^{MW}(k)$, which coincides with the one sketched in the introduction, for an actual morphism $\mathbb{P}^1 \to \mathbb{P}^1$ which has a regular value. However it is not an isomorphism in general: its kernel is isomorphic to the subgroup of squares $(k^\times)^2$ in $k^\times$.

Remark 4.12. 1) *Transfers.* It is well know that, given a finite separable field extension $k \subset L$ together with a primitive element $x \in L$ (which generates $L|k$), one can define a transfer morphism in $\mathsf{H}_\bullet(k)$ of the form

$$tr_x : \mathbb{P}^1 \to \mathbb{P}^1 \wedge (\mathrm{Spec}(L)_+).$$

This follows from the Purity Theorem of [38] (or the Thom–Pontryagin construction) applied to the closed immersion $\mathrm{Spec}(L) \to \mathbb{P}^1$ determined by x. Using our computations and methods, we have been able to show that the induced morphism on $\mathbb{H}_1^{\mathbb{A}^1}$ does not depend on the choice of x. As a consequence we obtain that for any strictly $\mathbb{A}^1$-invariant sheaf M the strictly $\mathbb{A}^1$-invariant sheaf M_{-1} has canonical transfers morphisms for finite separable extensions between separable extensions of k. This can be used to simplify the construction of transfers in Milnor K-theory [18], [7].

Beware however that this notion of transfers for finite extension is slightly more general than Voevodsky's notion. The sheaf M_{-1} is automatically a sheaf of modules

over $\underline{\mathbf{K}}_0^{MW}$. Given a finite separable extension $k \subset L$ as above, the composition $M_{-1}(k) \to M_{-1}(L) \xrightarrow{\mathrm{tr}} M_{-1}(k)$ is precisely the multiplication by the class of L in $K_0^{MW}(k)$ (which is its Euler characteristic by the remark below). In characteristic $\neq 2$, this is (up to an invertible element) the trace form of $L|k$ in the Grothendieck–Witt group. In the case of Voevodsky's structure this composition is just the multiplication by $[L : k] \in \mathbb{N}$.

2) Using the previous computations as well as the classical ideas on Atiyah duality [2] and [16] in $\mathbb{A}^1$-algebraic topology[5] one may define for any morphism f (in fact in $\mathsf{H}(k)$) from a smooth projective k-variety X to itself a Lefschetz number $\lambda(f) \in K_0^{MW}(k)$ which satisfies all the usual properties (like the Lefschetz fixed point formula). In particular the Euler characteristic of X lies in $K_0^{MW}(k)$.

3) In view of the cartesian diagram (4.2) and our philosophy, the part coming from the Milnor K-theory is the one compatible with the intuition coming from the topology of complex points (or motives), and the part coming from the Witt K-theory is the one compatible with the intuition on the topology of real points. For any $X \in \mathsf{Sm}_k$ the graded ring $\bigoplus_n H^n(X; \underline{\mathbf{K}}_n^{MW})$ maps surjectively to the Chow ring $CH^*(X) = \bigoplus_n H^n(X; \underline{\mathbf{K}}_n^M)$ and to the graded ring $\bigoplus_n H^n(X; \underline{\mathbf{K}}_n^W)$ (however it does not inject into the product in general: one has a Mayer–Vietoris type long exact sequence). Given a real embedding there exists a morphism of rings $\bigoplus_n H^n(X; \underline{\mathbf{K}}_n^W) \to H^*(X(\mathbb{R}); \mathbb{Z})$. Note that it is known that the Chow ring only maps to $H^*(X(\mathbb{R}); \mathbb{Z}/2)$.

5. Some results on classifying spaces in $\mathbb{A}^1$-homotopy theory

Serre's splitting principle and $\mathbb{H}_0^{\mathbb{A}^1}$ of some classifying spaces. The Serre's splitting principle was stated in [15] only in terms of étale cohomology groups §24 or in terms of Witt groups §29, but we may easily generalize it to our situation.

Let us briefly recall from [53] and also [38] the notion of *geometric classifying space* $B_{gm}(G)$ for a linear algebraic group G. Choose a closed immersion of k-groups $\rho\colon G \subset \mathbb{G}L_n$. For each $r > 0$, denote by $U_r \subset \mathbb{A}^{rn}$ the open subset where G acts freely (in the étale topology) in the direct sum of r copies of the representation ρ. $B_{gm}(G)$ is then the union over r of the quotient k-varieties U_r/G, which are smooth k-varieties. We proved in [38] that for G a finite group of order prime to $\mathrm{char}(k)$ and X a smooth k-variety:

$$\mathrm{Hom}_{\mathsf{H}(k)}(X, B_{gm}(G)) \cong H^1_{\text{ét}}(X; G).$$

For n an integer, denote by $m = [\frac{n}{2}]$ and by $(\mathbb{Z}/2)^m \subset \Sigma_n$ the natural embedding. The following result is a variation on the Splitting principle [15, §24] (using the fact

[5] These ideas are also present in much more elaborated form in Voevodsky formalism of cross-functors [59], see also [3].

that strictly $\mathbb{A}^1$-invariant sheaves have also residues [34] as well as [15, Appendix C, A letter from B. Totaro to J.-P. Serre]):

Theorem 5.1 (Serre's splitting principle). *For any strictly $\mathbb{A}^1$-invariant sheaf M the restriction map*

$$H^0(B_{gm}(\Sigma_n); M) \to H^0(B_{gm}((\mathbb{Z}/2)^m); M)$$

is injective.

Corollary 5.2. *The homomorphism*

$$\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}((\mathbb{Z}/2)^m) \to \mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(\Sigma_n))$$

is an epimorphism.

We observe that $B_{gm}((\mathbb{Z}/2)^m)$ is $\mathbb{A}^1$-equivalent to a point in characteristic 2, see [38]. In that case we get $\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(\Sigma_n)) = \mathbb{Z}$.

In characteristic $\neq 2$, one has an exact sequence $\tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \to \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \to \tilde{\mathbb{H}}_0^{\mathbb{A}^1}(B_{gm}(\mathbb{Z}/2)) \to 0$ where the left morphism is induced by the squaring map (this comes from the fact that $B_{gm}(\mathbb{Z}/2)$ is the union of the quotients $(\mathbb{A}^n - \{0\})/(\mathbb{Z}/2)$). Thus $\tilde{\mathbb{H}}_0^{\mathbb{A}^1}(B_{gm}(\mathbb{Z}/2)) = \underline{\mathbf{K}}_1^{MW}/h = \underline{\mathbf{K}}_1^W$ and $\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(\mathbb{Z}/2)) = \mathbb{Z} \oplus \underline{\mathbf{K}}_1^W$.

Now the $\mathbb{A}^1$-tensor product $\underline{\mathbf{K}}_n^W \otimes_{\mathbb{A}^1} \underline{\mathbf{K}}_m^W$ is $\underline{\mathbf{K}}_{n+m}^W$. Using this we may compute $\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}((\mathbb{Z}/2)^m))$ by the Künneth formula and as the morphism of Theorem 5.1 is invariant under the action of Σ_m we get in characteristic $\neq 2$ an epimorphism of sheaves

$$\bigoplus_{i\in\{0,\dots,m\}} \underline{\mathbf{K}}_i^W \twoheadrightarrow \mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(\Sigma_n)). \tag{5.1}$$

Theorem 5.3. *In characteristic $\neq 2$ the epimorphism* (5.1) *is an isomorphism.*

The method is to construct refined Stiefel–Whitney classes $W_i\colon K_0^{MW}(F) \to K_i^W(F)$ lifting the usual ones w_i in $k_i(F)$ using the same method as in [26, §3]. The composition $\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(\Sigma_n)) \to \mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(O_n)) \xrightarrow{\oplus W_i} \bigoplus_{i\in\{0,\dots,m\}} \underline{\mathbf{K}}_i^W$ is the required left inverse.

Remark 5.4. 1) This result implies the Baratt–Priddy–Quillen Theorem in dimension 0 (at least in characteristic $\neq 2$), stating that the morphism induced by the stable transfers

$$\amalg_{n\in\mathbb{N}} B_{gm}(\Sigma_n) \to Q_{\mathbb{P}^1}\mathbb{S}^0$$

where $Q_{\mathbb{P}^1}\mathbb{S}^0$ means the colimit of the iterated $\mathbb{P}^1$-loop spaces[6], is an $\mathbb{A}^1$-stable group completion[7], see [37].

[6] $\operatorname{colim}_n R\underline{\mathrm{Hom}}_\bullet((\mathbb{P}^1)^{\wedge n}, (\mathbb{P}^1)^{\wedge n})$

[7] Voevodsky proved that it is not the usual group completion.

2) The same computation holds for Suslin singular homology [52] of $B_{gm}(\Sigma_n)$: one gets in characteristic $\neq 2$: $\mathbb{H}_0^S(B_{gm}(\Sigma_n)) = \bigoplus_{i\in\{0,\dots,m\}} \underline{\mathbf{k}}_i$.

3) Using the refined Stiefel–Whitney classes W_i considered previously and [15] we can also compute in characteristic $\neq 2$: $\mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(O_n)) = \oplus_{i\in\{0,\dots,n\}} \underline{\mathbf{K}}_i^W$ and $\mathbb{H}_0^S(B_{gm}(O_n)) = \oplus_{i\in\{0,\dots,n\}} \underline{\mathbf{k}}_i$. We observe as a consequence that the natural map (of sets)

$$H^1_{\text{ét}}(k;\, O_n) \to \mathbb{H}_0^{\mathbb{A}^1}(B_{gm}(O_n))(k)$$

is injective (but is not if one consider the Suslin $\mathbb{H}_0^S$ instead !). It is a natural question to ask for which algebraic k-groups the analogous map is injective. It is wrong for finite groups in general (but the abelian ones). It could be however true for a general class of algebraic groups G, in connection with a conjecture of Serre addressing the injectivity of the extension map $H^1_{\text{ét}}(k; G) \to H^1_{\text{ét}}(L_1; G) \times H^1_{\text{ét}}(L_2; G)$ when the finite field extensions L_1 and L_2 have coprime degrees over k.

$\mathbb{A}^1$-homotopy classification of algebraic vector bundles. Lindel has proven in [25] that for any n and for any smooth affine k-scheme X the projection $X \times \mathbb{A}^1 \to X$ induces a bijection

$$H^1_{Zar}(X; \mathbb{G}L_n) \to H^1_{Zar}(X \times \mathbb{A}^1; \mathbb{G}L_n)$$

(after the fundamental cases obtained by Quillen [45] and Suslin [50] on the Serre problem). As a consequence if one denotes by $\mathbb{G}r_n$ the "infinite Grassmanian of n-plans" the natural map $\mathrm{Hom}_k(X; \mathbb{G}r_n) \to H^1_{Zar}(X; \mathbb{G}L_n)$ which to a morphism assigns the pull-back of the universal rank n bundle, induces a map $\pi(X; \mathbb{G}r_r) \to H^1_{Zar}(X; \mathbb{G}L_n)$ (where the source means the set of morphisms modulo naive $\mathbb{A}^1$-homotopies); it is moreover easy to show this map is a bijection.

Theorem 5.5 ([35]). *For any integer $n \geq 3$ and any affine smooth k-scheme X the obvious map*

$$H^1_{Zar}(X; \mathbb{G}L_n) \cong \pi(X; \mathbb{G}r_r) \to \mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{G}r_r)$$

is a bijection.

For $n = 1$ this is well-known [38]. The proof of this result relies on the works of Quillen, Suslin, Lindel cited above and also on the works of Suslin [51] and Vorst [60] on the generalized Serre problem for the general linear group. In these latter works n has to be assumed $\neq 2$. We conjecture however that the statement of the previous theorem should remain true also for $n = 2$.

One then observes that one has an $\mathbb{A}^1$-fibration sequence of pointed spaces:

$$\mathbb{A}^n - \{0\} \to \mathbb{G}r_{n-1} \to \mathbb{G}r_n \tag{5.2}$$

because the simplicial classifying space $B(\mathbb{G}L_m)$ is $\mathbb{A}^1$-equivalent to $\mathbb{G}r_m$, for any m, and because $\mathbb{G}L_n/\mathbb{G}L_{n-1} \to \mathbb{A}^n - \{0\}$ is an $\mathbb{A}^1$-weak equivalence. From Theorem 4.9

we know that the space $\mathbb{A}^n - \{0\}$ is $(n-2)$-connected and that there exists a canonical isomorphism of sheaves: $\pi_{n-1}^{\mathbb{A}^1}(\mathbb{A}^n - \{0\}) \cong \underline{\mathbf{K}}_n^{MW}$.

Euler class and Stably free vector bundles. For a given smooth affine k-scheme X and an integer $n \geq 4$ we may now study the map:

$$H^1_{Zar}(X; \mathbb{G}L_{n-1}) \to H^1_{Zar}(X; \mathbb{G}L_n)$$

of adding the trivial line bundle following the classical method of obstruction theory in homotopy theory:

Theorem 5.6 (Theory of Euler class, [35]). *Assume $n \geq 4$. Let X be a smooth affine k-scheme, together with an oriented algebraic vector bundle ξ of rank n (this means a vector bundle of rank n and a trivialization of $\Lambda^n(\xi)$). Define its Euler class*

$$e(\xi) \in H^n(X; \underline{\mathbf{K}}_n^{MW}) = H^n(X; \pi_{n-1}^{\mathbb{A}^1}(\mathbb{A}^n - \{0\}))$$

to be the obstruction class obtained from Theorem 5.5 *and the $\mathbb{A}^1$-fibration sequence* (5.2). *If dimension $X \leq n$ we have the following equivalence:*

$$\xi \textit{ split off a trivial line bundle} \Leftrightarrow e(\xi) = 0 \in H^n(X; \underline{\mathbf{K}}_n^{MW}).$$

Remark 5.7. 1) In case $\mathrm{char}(k) \neq 2$, the group $H^n(X; \underline{\mathbf{K}}_n^{MW})$ coincides with the oriented Chow group $\widetilde{CH}^n(X)$ as defined in [5] and our Euler class coincides also with the one defined in *loc. cit.* There is an epimorphism from the Euler class group of Nori [8] to ours but we do not know whether this is an isomorphism. We observe that in [8] an analogous result is proven, and our result implies the result in [8]. If $\mathrm{char}(k) \neq 2$, in [5] the case of rank $n = 2$ was settled by some other method.

2) If ξ is an algebraic vector bundle of rank n over X, let $\lambda_\xi = \Lambda^n(\xi) \in \mathrm{Pic}(X)$ denotes its first Chern class. The obstruction class $e(\xi)$ obtained by the $\mathbb{A}^1$-fibration sequence (5.2) lives now in the corresponding cohomology group $H^n(X; \underline{\mathbf{K}}_n^{MW}(\lambda_\xi))$ obtained by twisting the sheaf $\underline{\mathbf{K}}_n^{MW}$ by λ_ξ.

3) The obvious morphism

$$H^n(X; \underline{\mathbf{K}}_n^{MW}) \to H^n(X; \underline{\mathbf{K}}_n^{M}) = CH^n(X)$$

maps the Euler class to the top Chern class $c_n(\xi)$. When k is algebraically closed and $\dim(X) \leq n$, this homomorphism is an isomorphism. This case of the Theory is due to Murthy [39].

4) Given a real embedding of the base field $k \to \mathbb{R}$, the canonical morphism from Remark 4.12 3): $H^n(X; \underline{\mathbf{K}}_n^{MW}) \to H^n(X(\mathbb{R}); \mathbb{Z})$ maps the Euler class $e(\xi)$ to the Euler class of the real vector bundle $\xi(\mathbb{R})$.

The long exact sequence in homotopy for the $\mathbb{A}^1$-fibration sequence (5.2) (applied to $(n+1)$) also gives the following theorem (compare [9]):

Theorem 5.8 (Stably free vector bundles, [35]). *Assume $n \geq 3$. Let X be a smooth affine k-scheme. The canonical map*

$$\mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{A}^{n+1} - \{0\}) \,/\, \mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{G}L_{n+1}) \to \mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{G}r_n)$$

is injective and its image $\Psi_n(X) \subset H^1_{Zar}(X; \mathbb{G}L_n) = \mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{G}r_n)$ consists exactly of the set of isomorphism classes of algebraic vector bundles of rank n over X such that $\xi \oplus \theta^1$ is trivial.

Moreover if the dimension of X is $\leq n$, the natural map

$$\mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{A}^{n+1} - \{0\}) \to H^n(X; \underline{\mathbf{K}}^{MW}_{n+1})$$

is a bijection and the natural action of $\mathrm{Hom}_{\mathsf{H}(k)}(X, \mathbb{G}L_{n+1})$ factors trough the determinant as an action of $\mathcal{O}(X)^\times$. In that case, we get a bijection

$$H^n(X; \underline{\mathbf{K}}^{MW}_{n+1})/\mathcal{O}(X)^\times \cong \Psi_n(X).$$

Remark 5.9. Using Popescu's approximation result [43] it is possible, with some care, to extend the results of this paragraph to affine regular schemes defined over a field k.

6. Miscellaneous

Proofs of the Milnor conjecture on quadratic forms. Using Voevodsky's result [57] we have produced two proofs of the Milnor conjecture on quadratic forms asserting that for a field F of characteristic $\neq 2$ the Milnor epimorphism $s_F : k_*(F) \to i^*(F)$ is an isomorphism.

The first one is only sketched in [29], however it is very striking in the context of $\mathbb{A}^1$-algebraic topology. We consider the Adams spectral sequence based on mod 2 motivic cohomology "converging" to $\pi_*^{\mathbb{A}^1}(\mathbb{S}^0)$. Using an unpublished work of Voevodsky on the computation of the mod 2 motivic Steenrod algebra we showed that $E_2^{s,u} = \mathrm{Ext}^s_{\mathcal{A}_k}(H^*(k; \mathbb{Z}/2(*)), H^*(k; \mathbb{Z}/2(*))[s+u])$ and could compute enough. First $E_2^{s,u}$ vanishes for $u < 0$ which is compatible with the $\mathbb{A}^1$-connectivity result 3.5, which implies $\pi_u^{\mathbb{A}^1}(\mathbb{S}^0) = 0$ for $u < 0$. More striking is the computation of the column $E_2^{s,0}$ converging to $\pi_0^{\mathbb{A}^1}(\mathbb{S}^0) = GW(k)$ (in characteristic $\neq 2$). We found that $E_2^{0,0} = \mathbb{Z}/2$ and that for $s > 0$

$$E^{s,0} = \mathbb{Z}/2 \oplus k_s(k).$$

This is exactly the predicted form of the associated graded ring for $GW(k)$ by the Milnor conjecture. The terms $\mathbb{Z}/2$ are detected (in the bar complex) by the tensor powers of the Bockstein $\beta^{\otimes s}$ and the mod 2 Milnor K-theory terms are detected by the tensor powers of the $\mathbb{S}q^2$-operation[8] of Voevodsky $(\mathbb{S}q^2)^{\otimes s}$. The proof of the Milnor

[8] This relationship is explained again by "taking" the real points: the operation $\mathbb{S}q^2$ "induces" the Bockstein operation on mod 2 singular cohomology of real points

conjecture then amounts to showing that the Adams spectral sequence degenerates from the $E_2^{*,*}$-term on the column $u = 0$.

The degenerescence was obtained by a careful study of the column $E_2^{*,1}$ from which the potential differentials start to reach $E_2^{*,0}$, using the Milnor conjecture on mod 2 Galois cohomology of fields of characteristic 2 established by Voevodsky in [57]. The idea was to observe that the groups $E_2^{*,1}$ are enough "divisible" by some suitable mod 2-Milnor K-theory groups. We realized recently in [33] that this argument could be made much simpler and that everything amounts to proving some "$\mathbb{P}^1$-cellularity" of the sheaves $\underline{\mathbf{k}}_n$ in the $\mathbb{A}^1$-derived category, which again is given by the main result of [57].

Global properties of the stable $\mathbb{A}^1$-homotopy category. We have unfortunately no room available to discuss much recent developments in the global properties of the stable $\mathbb{A}^1$-homotopy category. Let us just mention briefly: our work (in preparation) on the rational stable homotopy category and its close relationship with Voevodsky's category of rational mixed motives. The slice filtration and motivic Atiyah–Hirzebruch's type spectral sequence approach due to Voevodsky (see [58] for instance); we must also mention Levine's recent work in this direction, for instance [22]. There is also a work in preparation by Hopkins and the author starting from the Thom spectrum $M\mathbb{G}L$, where is proven that the "homotopical quotient" $M\mathbb{G}L/(x_1, \dots, x_n, \dots)$ obtained by killing the generators of the Lazard ring is, in characteristic 0, the motivic cohomology spectrum of Voevodsky. This gives an Atiyah–Hirzebruch spectral sequence for $M\mathbb{G}L$ (and also K-theory) and gives an other (purely homotopical) proof of the general degree formula of [24], [23].

We must mention Voevodsky's formalism of cross functors [59] and Ayoub's work [3] in which is established the analogue of the theory of vanishing cycles in the context of Voevodsky's triangulated category of motives.

From $\mathbb{A}^1$-homotopy to algebraic geometry? We conclude this paper by an observation. All the tools and notions concerning the classical approach to surgery in classical differential topology seem now available in $\mathbb{A}^1$-algebraic topology: degree, homology, fundamental groups, cobordism groups [24], [23], Poincaré complexes, classification of vector bundles, etc. We also have natural candidates of surgery groups using Balmer's Witt groups [4] of some triangulated category of $\pi_1^{\mathbb{A}^1}$-modules. Why not then dreaming about a surgery approach also for smooth projective k-varieties? Of course there is no obvious analogues for surgery. There is also a major new difficulty: we have observed that even the simplest varieties like the projective spaces are never simply connected. This fact obstructs any hope of "h-cobordism" theorem[9], but now we also understand the reason: the $\mathbb{A}^1$-fundamental group of a pointed projective smooth k-scheme is almost never trivial. A major advance would then be to find the analogue of the "s-cobordism" theorem, the generalization of the h-cobordism theorem in the presence of π_1.

[9] Marc Levine indeed produced a counter-example

References

[1] Arason, J. K., and Elman, R., Powers of the fundamental ideal in the Witt ring. *J. Algebra* **239** (2001), 150–160.

[2] Atiyah, M. F., Thom complexes. *Proc. London Math. Soc.* (3) **11** (1961), 291–310.

[3] Ayoub, J., Les six opérations de Grothendieck et le formalisme des cycles evanescents dans le monde motivique. Thèse de l'université Paris 7, 2006; http:// www.math.uiuc.edu/K-theory/0761/.

[4] Balmer, P., An introduction to triangular Witt groups and a survey of applications. In *Algebraic and arithmetic theory of quadratic forms* (Talca, Chile, 2002), Contemp. Math. 344, Amer. Math. Soc., Providence, RI, 2004, 31–58.

[5] Barge, J., et Morel, F., Cohomologie des groupes linéaires. K-théorie de Milnor et groupes de Witt. *C. R. Acad. Sci. Paris Sér. I Math.* **328** (1999), 191–196.

[6] Barge, J., et Morel, F., Groupe de Chow des cycles orientés et classe d'Euler des fibrés vectoriels. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (2000), 287–290.

[7] Bass, H., Tate, J., The Milnor ring of a global field. In *Algebraic K-theory, II: "Classical" algebraic K-theory and connections with arithmetic* (Proc. Conf., Seattle, Wash., Battelle Memorial Inst., 1972), Lecture Notes in Math. 342, Springer-Verlag, Berlin 1973, 349–446.

[8] Bhatwadekar, S. M., Sridharan, R., The Euler class group of a Noetherian ring. *Compositio Math.* **122** (2) (2000), 183–222.

[9] Bhatwadekar, S. M., Sridharan, R., On Euler classes and stably free projective modules. In *Algebra, arithmetic and geometry* (Mumbai, 2000), Part I, Tata Inst. Fund. Res. Stud. Math. 16, Tata Inst. Fund. Res., Bombay 2002, 139–158.

[10] Brown, K. S., Abstract homotopy theory and generalized sheaf cohomology. *Trans. Amer. Math. Soc.* **186** (1973), 419–458.

[11] J.-L. Colliot-Thélène, R. T. Hoobler, B. Kahn, The Bloch-Ogus-Gabber theorem. In *Algebraic K-theory* (Toronto, ON, 1996), Fields Inst. Commun. 16, Amer. Math. Soc., Providence, RI, 1997, 31–94.

[12] Déglise, F., Modules homotopiques avec transferts et motifs génériques. Thèse de l'université Paris 7, 2002; http://www-math.univ-paris13.fr/ deglise/these.html

[13] Deligne, P., and Brylinski, J. L., Central extensions of reductive groups by $\underline{\mathbf{K}}_2$. *Inst. Hautes Études Sci. Publ. Math.* **94** (2001), 5–85.

[14] Gabber, O., Gersten's conjecture for some complexes of vanishing cycles. *Manuscripta Math.* **85** (3–4) (1994), 323–343.

[15] Garibaldi, S., Merkurjev, A., Serre, J.-P., *Cohomological Invariants in Galois Cohomology*. Amer. Math. Soc. Univ. Lecture Ser. 28, Amer. Math. Soc., Providence, RI, 2003.

[16] Hu, P., On the Picard group of the stable $\mathbb{A}^1$-homotopy category. *Topology* **44** (3) (2005), 609–640.

[17] Hu, P., and Kriz, I., The Steinberg relation in $\mathbb{A}^1$-stable homotopy. *Internat. Math. Res. Notices* **2001** (17) (2001), 907–912.

[18] Kato, K., A generalization of local class field theory by using K-groups. II. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **27** (3) (1980), 603–683.

[19] Kato, K., Milnor K-theory and the Chow group of zero cycles. In *Applications of algebraic K-theory to algebraic geometry and number theory* (Boulder, Colo., 1983), Part I, Contemp. Math. 55, Amer. Math. Soc., Providence, RI, 1986, 241–253.

[20] Kato, K., Symmetric bilinear forms, quadratic forms and Milnor K-theory in characteristic two. *Invent. Math.* **66** (3) (1982), 493–510.

[21] Jardine, J. F., Simplicial presheaves. *J. Pure Appl. Algebra* **47** (1) (1987), 35–87.

[22] Levine, M., The homotopy coniveau filtration. Preprint; http://www.math.neu.edu/.

[23] Levine, M., Morel, F., Cobordisme algébrique I & II. *C. R. Acad. Sci. Paris Sér. I Math.* **332** (2001) 723–728; 815–820.

[24] Levine, M., Morel, F., Algebraic cobordism. To appear.

[25] Lindel, H., On the Bass-Quillen conjecture concerning projective modules over polynomial rings. *Invent. Math.* **65** (2) (1981/82), 319–323.

[26] Milnor, J., Algebraic K-theory and quadratic forms. *Invent. Math.* **9** (1969/1970), 318–344.

[27] Milnor, J., Husemoller, D., *Symmetric bilinear forms*. Ergeb. Math. Grenzgeb. 73, Springer-Verlag, New York, Heidelberg 1973.

[28] Morel, F., Théorie homotopique des schémas. *Astérisque* **256** (1999).

[29] Morel, F., Suite spectrale d'Adams et invariants cohomologiques des formes quadratiques. *C. R. Acad. Sci. Paris Sér. I Math.* **328** (1999), 963–968.

[30] Morel, F., An introduction to A^1-homotopy theory. In *Contemporary Developments in Algebraic K-theory* (ed. by M. Karoubi, A. O. Kuku , C. Pedrini), ICTP Lecture Notes 15, Abdus Salam Int. Cent. Theoret. Phys., Trieste 2004, 357–441.

[31] Morel, F., On the motivic stable π_0 of the sphere spectrum. In *Axiomatic, Enriched and Motivic Homotopy Theory* (ed. by J. P. C. Greenlees), Kluwer Academic Publishers, Dordrecht 2004, 219–260.

[32] Morel, F., Sur les puissances de l'idéal fondamental de l'anneau de Witt. *Comment. Math. Helv.* **79** (4) (2004), 689–703.

[33] Morel, F., Milnor's conjecture on quadratic forms and mod 2 motivic complexes. *Rend. Sem. Mat. Univ. Padova* **114** (2005), 51–62 .

[34] Morel, F., The stable $\mathbb{A}^1$-connectivity theorems. *K-theory*, to appear.

[35] Morel, F., $\mathbb{A}^1$-homotopy classification of vector bundles over smooth affine schemes. Preprint; http://www.mathematik.uni-muenchen.de/~morel/listepublications.html

[36] Morel, F., On the structure of $\mathbb{A}^1$-homotopy sheaves. Preprint; http:// www.mathematik.uni-muenchen.de/~morel/listepublications.html

[37] Morel, F., Serre's splitting principle and the motivic Barrat-Priddy-Quillen theorem. In preparation.

[38] Morel, F., Voevodksy, V., $\mathbb{A}^1$-homotopy theory of schemes. *Inst. Hautes Études Sci. Publ. Math.* **90** (1999), 45–143.

[39] Murthy, M.P., Zero cycles and projective modules. *Ann. of Math.* **140** (1994), 405–434.

[40] Nisnevich, Y. A., The completely decomposed topology on schemes and associated descent spectral sequences in algebraic K-theory. In *Algebraic K-theory: connections with geometry and topology* (Lake Louise, AB, 1987), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 279, Kluwer Academic Publishers, Dordrecht 1989, 241–342.

[41] Orlov, D., Vishik, A., Voevodsky, V., An exact sequence for Milnor's K-theory with applications to quadratic forms. Preprint; http://www.math.uiuc.edu/K-theory/0454/.

[42] Panin, I., Homotopy invariance of the sheaf WNis and of its cohomology. Preprint; http://www.math.uiuc.edu/K-theory/0715/.

[43] Popescu, D., Néron desingularisation and approximation. *Nagoya Math. J.* **104** (1986), 85–115.

[44] Quillen, D., *Homotopical algebra*. Lecture Notes in Math. 43, Springer-Verlag, Berlin 1967.

[45] Quillen, D., Projective modules over polynomial rings. *Invent. Math.* **36** (1976), 167–171.

[46] Rost, M., Chow groups with coefficients. *Doc. Math.* **1** (16) (1996), 319–393 (electronic).

[47] Scharlau, W., *Quadratic and Hermitian forms*. Grundlehren Math. Wiss. 270, Springer-Verlag, Berlin 1985.

[48] Schmid, M., Wittringhomologie. Dissertation, Fakultät für Mathematik der Universität Regendburg, 1998.

[49] Suslin, A., Torsion in K_2 of fields. *K-Theory* **1** (1) (1987), 5–29.

[50] Suslin, A., Projective modules over polynomial rings. *Mat. Sb.* (*N.S.*) **93** (135) (1974), 588–595, 630 (in Russian).

[51] Suslin, A., The structure of the special linear group over rings of polynomials. *Izv. Akad. Nauk SSSR Ser. Mat.* **41** (2) (1977), 235–252, 477 (in Russian).

[52] Suslin, A., Voevodsky, V., Singular homology of abstract algebraic varieties. *Invent. Math.* **123** (1) (1996), 61–94.

[53] Totaro, B., The Chow ring of a classifying space. In *Algebraic K-theory* (Seattle, WA, 1997), Proc. Sympos. Pure Math. 67, Amer. Math. Soc., Providence, RI, 1999, 249–281.

[54] Voevodksy, V., $\mathbf{A}^1$-homotopy theory. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. I, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 579–604.

[55] Voevodksy, V., Cohomological theory of presheaves with transfers. In *Cycles, transfers, and motivic homology theories*, Ann. of Math. Stud. 143, Princeton University Press, Princeton, NJ, 2000, 87–137.

[56] Voevodksy, V., Triangulated categories of motives over a field. In *Cycles, transfers, and motivic homology theories*, Ann. of Math. Stud. 143, Princeton University Press, Princeton, NJ, 2000, 188–238.

[57] Voevodsky, V., Motivic cohomology with $Z/2$-coefficients. *Publ. Math. Inst. Hautes Études Sci.* **98** (2003), 59–104.

[58] Voevodsky, V., On the zero slice of the sphere spectrum. Preprint; http://www.math.uiuc.edu/K-theory/0612/

[59] Voevodsky, V., "Voevodsky lectures on cross functors", by P. Deligne. Preprint; http://www.math.ias.edu/ vladimir/delnotes01.ps

[60] Vorst, The Serre problem for discrete Hodge algebra. *Math. Z.* **184** (1983), 425–433.

Mathematisches Institut der Universität München, Theresienstr. 39, 80333 München, Germany

E-mail: morel@mathematik.uni-muenchen.de

Development in symplectic Floer theory

Kaoru Ono*

Abstract. In the middle of the 1980s, Floer initiated a new theory, which is now called the Floer theory. Since then the theory has been developed in various ways. In this article we report some recent progress in Floer theory in symplectic geometry. For example, we give an outline of a proof of the flux conjecture, which states that the Hamiltonian diffeomorphism group is C^1-closed in the group of symplectomorphisms for closed symplectic manifolds. We also give a brief survey on the obstruction–deformation theory for Floer theory of Lagrangian submanifolds and explain some of its applications.

Mathematics Subject Classification (2000). Primary 53D40; Secondary 53D35.

Keywords. Symplectic manifold, Hamiltonian systems, Lagrangian submanifolds, Floer cohomology, A_∞-structure.

1. Introduction

In [9]–[14], Andreas Floer initiated "$\frac{\infty}{2}$-dimensional" (co)homology theory, which is now called Floer theory. He invented this theory to prove Arnold's conjecture for fixed points of Hamiltonian diffeomorphism and, under certain assumptions, its analogue for Lagrangian intersections. Roughly speaking, the conjecture states that there is a non-trivial topological lower bound for the number of fixed points of a Hamiltonian diffeomorphism. It is one of his conjectures which stimulated recent developments in symplectic geometry. This theory was soon adapted in Donaldson theory and he constructed the instanton homology theory. A lot of work has been done since and Floer theory has been developed in various directions. In this article, we will describe some recent development of Floer theory in symplectic geometry.

In these decades, symplectic geometry has been much developed. In particular, Gromov revealed many significant phenomena based on his theory of pseudo-holomorphic curves [18] and revolutionized the study in this area. Hamiltonian dynamics is one of main sources of symplectic geometry. The existence of periodic trajectories is a basic problem and there are many works on this subject up to now. In fact, the existence of periodic trajectories reflects so-called symplectic rigidity phenomena. Since trajectories of a Hamiltonian system are characterized by the least action principle, the variational method can be applied to the existence of periodic

*The author is partly supported by Grant-in-Aid for Scientific Research Nos. 14340019 and 17654009, Japan Society for the Promotion of Science.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

trajectories. Namely, closed trajectories are critical points of the action functional associated to the Hamiltonian system. Floer combined the variational framework with the theory of pseudo-holomorphic curves to construct an analogue of Morse theory for the action functional.

In the first part of this article, we discuss Floer theory for Hamiltonian systems and present some applications including the flux conjecture. In the second part, we discuss Floer theory for Lagrangian submanifolds. In general, Floer cohomology may not be defined for a pair of Lagrangian submanifolds. We briefly describe the obstruction to defining Floer cohomology as well as the filtered A_∞-algebra associated to a Lagrangian submanifold. We also present some applications, e.g., Lagrangian intersections, non-triviality of the Maslov class, etc. Although there will be some overlaps with Y. G. Oh's contribution to this proceedings, we will try to put different emphases on the theory in this lecture.

2. Floer theory for symplectomorphisms

2.1. Review on the construction. In this section, we briefly review the construction of Floer cohomology for symplectomorphisms, especially Hamiltonian diffeomorphisms, which was initiated in [14] and developed in e.g., [21], [36], [16], [29]. Let (M, ω) be a symplectic manifold. In this article, we assume that M is compact without boundary for simplicity. Denote by X_h the Hamiltonian vector field of h defined by

$$i(X_h)\omega = dh.$$

For $H = \{h_t\}_{t\in\mathbb{R}}$, we integrate the time-dependent vector field X_{h_t} to obtain the one-parameter family $\{\varphi_t^H\}$ of diffeomorphisms. We call such $\{\varphi_t^H\}$ a time-dependent Hamiltonian flow. A diffeomorphism φ of M is called a Hamiltonian diffeomorphism, when φ is the time-one map of $\{\varphi_t^H\}$ for some H. We may assume that $h_{t+1} = h_t$. Denote by $\mathrm{Ham}(M, \omega)$ resp. $\mathrm{Symp}(M, \omega)$ the group of Hamiltonian diffeomorphisms resp. the group of symplectomorphisms which are diffeomorphisms preserving ω. Clearly, $\mathrm{Ham}(M, \omega) \subset \mathrm{Symp}(M, \omega)$. Hamiltonian diffeomorphisms are fundamental in symplectic geometry and enjoy some distinguished properties, e.g., existence of fixed points (see Arnold's conjecture below), simplicity [1], existence of a biinvariant distance on (the universal covering group of) $\mathrm{Ham}(M, \omega)$, called Hofer's distance, etc. Now we recall the following:

Conjecture 2.1 (Arnold's conjecture). For $\varphi \in \mathrm{Ham}(M, \omega)$ there are as many fixed points of φ as the smallest number of critical points of smooth functions on M, namely,

$$\#\,\mathrm{Fix}(\varphi) \geq \min\{\#\mathrm{Crit}(f) \mid f \in C^\infty(M)\}.$$

If all the fixed points of φ are non-degenerate, i.e. 1 is not an eigenvalue of $d\varphi$ at any

fixed point, then

$$\#\,\mathrm{Fix}(\varphi) \geq \min\{\#\mathrm{Crit}(f) \mid f \text{ is a Morse function on } M\}.$$

This conjecture has been verified for closed oriented surfaces, the torus, complex projective spaces, etc. A weaker version of the conjecture is formulated by replacing the lower bounds with the cup-length and the sum of Betti numbers, respectively. We call it homological Arnold conjecture.

Let φ be a symplectomorphism of (M, ω) such that all fixed points are non-degenerate. Following [6], we introduce the twisted loop space

$$\mathcal{P}_\varphi = \{\sigma : [0, 1] \to M \mid \varphi(\sigma(1)) = \sigma(0)\},$$

and define a closed 1-form, in a formal sense, on $\mathcal{P}_\varphi$ by

$$\alpha_\varphi(\xi) = \int_0^1 \omega(\xi, \dot{\sigma})\, dt \quad \text{for } \xi \in T_\sigma \mathcal{P}_\varphi.$$

Clearly, fixed points of φ are in one-to-one correspondence with zeros of α_φ. We take the smallest covering space $\pi : \widetilde{\mathcal{P}}_\varphi \to \mathcal{P}_\varphi$ such that (1) $\pi^*\alpha_\varphi$ is exact, i.e., there exists a primitive function $\mathcal{A}_\varphi$ for α_φ, and (2) the integer valued Maslov index μ is well defined on $\mathrm{Crit}(\mathcal{A}_\varphi) = \pi^{-1}(\mathrm{Zero}(\alpha_\varphi))$. From now on we call such a covering space the Floer covering space. Pick an almost complex structure $J = \{J_t\}$ compatible with ω such that $\varphi_* J_1 = J_0$. Then the gradient of $\mathcal{A}_\varphi$ is formally written as

$$\mathrm{grad}\, \mathcal{A}_\varphi(\sigma) = -J\dot{\sigma},$$

and gradient flow lines are regarded as solutions of the following equation

$$\frac{\partial u}{\partial \tau} + J_t(u)\frac{\partial u}{\partial t} = 0$$

for

$$u = u(\tau, t) : \mathbb{R} \times [0, 1] \to M \quad \text{such that } \varphi(u(\tau, 1)) = u(\tau, 0).$$

We set

$$CF^*(\varphi, J) = \Big\{ \sum_i a_i \tilde{\sigma}_i \mid a_i \in \mathbb{Q},\ \tilde{\sigma}_i \in \mathrm{Crit}(\mathcal{A}_\varphi) \text{ satisfy the following condition:}$$
$$\#\{i \mid a_i \neq 0,\ \mathcal{A}_\varphi(\tilde{\gamma}_i) < c\} \text{ is finite for any } c \in \mathbb{R}\Big\}.$$

The grading is given by the Maslov index μ on $\mathrm{Crit}(\mathcal{A}_\varphi)$. The coboundary operator $\delta = \delta^{\varphi, J}$ is defined by counting gradient flow lines connecting the critical points $\tilde{\sigma}^\pm$ of $\mathcal{A}_\varphi$ such that $\mu(\tilde{\sigma}^+) - \mu(\tilde{\sigma}^-) = 1$. Note that the covering transformation group $G_{M,\varphi}$ of $\pi : \widetilde{\mathcal{P}}_\varphi \to \mathcal{P}_\varphi$ naturally acts on the Floer complex. In fact, this action

extends to the so-called Novikov ring associated to $\varphi \in \mathrm{Symp}(M, \omega)$, which is a certain completion of the group ring of $G_{M,\varphi}$. To make this construction rigorous, we need to study compactness properties, transversality, etc. for the moduli space of solutions of the J-holomorphic curve equation above. We can achieve these points as in [16], [29], see also [28], [41], [45] based on the notion of stable maps [23], [24]. The resulting cohomology is the Floer cohomology $HF^*(\varphi, J)$, which is a module over the Novikov ring associated to $\varphi \in \mathrm{Symp}(M, \omega)$. We also find that Floer cohomology is invariant under Hamiltonian deformations of φ. In the case that M is 2-dimensional, Seidel noticed that the Floer cohomology is invariant under a class of deformations which contains all Hamiltonian deformations [43].

When $\varphi \in \mathrm{Symp}_0(M, \omega)$, i.e., there is a path φ_t, $0 \le t \le 1$, such that $\varphi_0 = \mathrm{id}$ and $\varphi_1 = \varphi$, we can formulate the Floer theory on the loop space LM of M rather than the twisted loop space $\mathcal{P}_\varphi$. (From now on, we call φ_t with $\varphi_0 = \mathrm{id}$ a based path.) Namely, we identify them by

$$\sigma(t) \in \mathcal{P}_\varphi \mapsto \gamma(t) = (\varphi_t)^{-1}(\sigma(t)) \in LM.$$

In particular, when $\varphi \in \mathrm{Ham}(M, \omega)$ we choose a based path φ_t in $\mathrm{Ham}(M, \omega)$. Denote by H the time-dependent Hamiltonian function which generates φ_t. Then fixed points of φ are in one-to-one correspondence with 1-periodic orbits of the time-dependent flow φ_t, which are characterized as zeros of the following closed 1-form α_H on the loop space LM of M:

$$\alpha_H(\xi) = \int_0^1 \omega(\xi(t), \dot{\gamma}(t) - X_{H_t}(\gamma(t)))\, dt,$$

where $\gamma \in LM$ and $\xi \in T_\gamma LM$, i.e., a section of $\gamma^* TM$. Write $J_t' = (\varphi_t)_* J_t$. (Note that $J_0' = J_1'$.) Then gradient flow lines are solutions of the following equation:

$$\frac{\partial u}{\partial \tau} + J'(u)\left(\frac{\partial u}{\partial t} - X_{H_t}(u)\right) = 0,$$

for $u = u(\tau, t) \colon \mathbb{R} \times S^1 \to M$. Denote by $p \colon \tilde{L}M \to LM$ the Floer covering space of LM and by $\mathcal{A}_H \colon \tilde{L}M \to \mathbb{R}$ the action functional, i.e., $d\mathcal{A}_H = p^* \alpha_H$. Consider the graded module generated by the critical points of $\mathcal{A}_H$ with the grading given by μ, which is known as the Conley–Zehnder index in this setting. Then take its completion with respect to the filtration $\{\tilde{\gamma} \in \mathrm{Crit}(\mathcal{A}_H) \mid \mathcal{A}_H(\tilde{\gamma}) > c\}$ for $c \in \mathbb{R}$. We denote it by $CF^*(H)$. The coboundary operator $\delta = \delta^{H,J}$ is defined by counting the number of connecting orbits joining critical points. In this case Floer cohomology can be computed as follows:

$$HF^*(H, J) \cong H^{*+n}(M; \mathbb{Q}) \otimes \Lambda_\omega,$$

where Λ_ω is the Novikov ring of (M, ω).

As a corollary we have the following result ([16], [29]).

Theorem 2.2. *If $\varphi \in \mathrm{Ham}(M, \omega)$ has only non-degenerate fixed points then*

$$\#\,\mathrm{Fix}(\varphi) \geq \sum_k \mathrm{rank}\ H^k(M; \mathbb{Q}).$$

More precisely, we can find that the number of fixed points which correspond to contractible 1-periodic orbits of any based Hamiltonian path is at least the sum of Betti numbers in this theorem. As a consequence, we also find that there always exists a contractible 1-periodic orbit for any time-dependent periodic Hamiltonian system. For a based path $\{\psi_t\}$ in $\mathrm{Symp}_0(M, \omega)$ we can define the Floer cohomology, which we may call the Floer–Novikov cohomology, in a way similar to the case of Hamiltonian diffeomorphisms. Under the $\pm$-monotonicity assumption we have a similar computation for $\varphi \in \mathrm{Symp}_0(M, \omega)$ using Novikov cohomology of the flux of φ_t in place of the ordinary cohomology of M, see [27].

2.2. Application to the flux conjecture. The flux of a based path $\{\phi_t\}_{0\leq t\leq 1}$ in $\mathrm{Symp}_0(M, \omega)$ is defined to be

$$\widetilde{\mathrm{Flux}}(\phi_t) = \int_0^1 [i(X_t)\omega]\, dt \in H^1(M; \mathbb{R}),$$

where X_t is the family of symplectic vector fields generating ϕ_t. The flux depends only on the homotopy class of paths with fixed end points $\phi_0 = \mathrm{id}$ and ϕ_1, and induces a homomorphism from the universal covering group $\widetilde{\mathrm{Symp}}_0(M, \omega)$ of $\mathrm{Symp}_0(M, \omega)$ to $H^1(M; \mathbb{R})$. Denote by Γ_ω, which is called the flux group, the image of $\mathrm{Ker}\left(\widetilde{\mathrm{Symp}}_0(M; \omega) \to \mathrm{Symp}_0(M; \mathbb{R})\right) \cong \pi_1(\mathrm{Symp}_0(M; \mathbb{R}))$ under $\widetilde{\mathrm{Flux}}$. It is known that the path ϕ_t above can be homotoped to a path in $\mathrm{Ham}(M; \omega)$ keeping the end points fixed if and only if $\widetilde{\mathrm{Flux}}(\phi_t) = 0$. $\widetilde{\mathrm{Flux}}$ descends to a homomorphism $\mathrm{Flux}\colon \mathrm{Symp}_0(M; \omega) \to H^1(M; \mathbb{R})/\Gamma_\omega$. The group $\mathrm{Ham}(M; \mathbb{R})$ is also known to be the kernel of this homomorphism, see [1]. Hence, it is a basic question in order to understand $\mathrm{Ham}(M, \omega) \subset \mathrm{Symp}_0(M, \omega)$ how Γ_ω is embedded in $H^1(M; \mathbb{R})$.

Conjecture 2.3 (Flux conjecture). Γ_ω is discrete in $H^1(M; \mathbb{R})$.

This conjecture is equivalent to that $\mathrm{Ham}(M, \omega)$ is C^1-closed in $\mathrm{Symp}_0(M, \omega)$. There are various cases in which the flux conjecture is verified. For example, if $[\omega] \in H^1(M; \mathbb{Q})$ the conjecture clearly holds. A less trivial case is that (M, ω) is of Lefschetz type, i.e., $\wedge[\omega]^{n-1}\colon H^1(M; \mathbb{R}) \to H^{2n-1}(M; \mathbb{R})$ is an isomorphism. It was Lalonde, McDuff and Polterovich [25], [26] who noticed that the affirmative answer to the homological Arnold conjecture can be used to prove the flux conjecture. Among other things, they proved the following:

Theorem 2.4. *If $c_1(M)\colon \pi_2(M) \to \mathbb{Z}$ is trivial or its minimal positive value (the minimal Chern number) is at least $2n = \dim_{\mathbb{R}} M$, then the flux conjecture holds.*

Theorem 2.5. *The rank of the flux group Γ_ω is at most the first Betti number $b_1(M)$ of M. In particular, the flux conjecture holds if $b_1(M) = 1$.*

Remark 2.6. Note that Theorem 2.5 follows from the flux conjecture.

We give an outline of the proof of the flux conjecture. First of all, we collect some notation and fundamental properties of the Floer–Novikov cohomology. For any based path $\{\psi_t\}$ in $\mathrm{Symp}_0(M, \omega)$, we can deform it by a homotopy so that $i(X_t)\omega$ does not depend on t and is equal to $\theta = \widetilde{\mathrm{Flux}}(\psi_t)$. Here X_t is the family of symplectic vector fields generating ψ_t. Denote by $\pi : \overline{M} \to M$ the covering space of M associated to the homomorphism $I_\theta : \pi_1(M) \to \mathbb{R}$ obtained by integrating θ along loops. Then there exists $\widetilde{H} = \{\tilde{h}_t\}$, a smooth family of smooth functions on $\overline{M}$ such that $\pi^* i(X_t)\omega = d\tilde{h}_t$. Denote by $\tilde{L}^\theta M$ the Floer covering space of LM for $\{\psi_t\}$ which depends only on its flux θ, and by $G_{\omega,\theta}$ its covering transformation group. Then we can perform the Floer construction for $\mathcal{A}_{\widetilde{H}} : \tilde{L}^\theta M \to \mathbb{R}$ and obtain the cochain complex $(CFN^*(\widetilde{H}, J), \delta = \delta^{\widetilde{H},J})$. The group $G_{\omega,\theta}$ naturally acts on this complex. Moreover, the action extends to the Novikov completion $\Lambda_{\omega,\theta}$ of the group ring of $G_{\omega,\theta}$. Denote by $HFN^*(\{\psi_t\})$ the resulting cohomology, which is the Floer–Novikov cohomology of $\{\psi_t\}$ and which is a finitely generated module over $\Lambda_{\omega,\theta}$.

We collect its fundamental properties as follows.

Theorem 2.7. *For based paths $\{\psi_t^{(1)}\}$ and $\{\psi_t^{(2)}\}$ with $\widetilde{\mathrm{Flux}}(\{\psi_t^{(1)}\}) = \widetilde{\mathrm{Flux}}(\{\psi_t^{(2)}\})$ we have a natural isomorphism*

$$HFN^*(\{\psi_t^{(1)}\}) \cong HFN^*(\{\psi_t^{(2)}\}).$$

Theorem 2.8. *If $\widetilde{\mathrm{Flux}}(\{\psi_t\})$ is sufficiently small we have*

$$HFN^*(\{\psi_t\}) \cong HN^{*+n}(\theta) \otimes_{\overline{\Lambda}_\theta} \Lambda_{\omega,\theta}.$$

Here $HN^*(\theta)$ is the Novikov cohomology of θ and $\overline{\Lambda}_\theta$ is its coefficient ring.

Secondly, we note that the Floer construction can be performed with coefficients in a local system as in the ordinary cohomology theory, see e.g., [38], [39]. In particular, when the flux vanishes, i.e. $\{\psi_t\}$ is a Hamiltonian path, we obtain the Floer cohomology for based Hamiltonian paths with coefficients in a local system. Let $L \to M$ be a local system or a flat vector bundle. We denote by $HFN^*(\{\psi_t\}; L)$ the Floer–Novikov cohomology of $\{\psi_t\}$ with coefficients in L. Then Theorems 2.7 and 2.8 holds with coefficients in L. We state them for reference.

Theorem 2.9. *Let $L \to M$ be a flat vector bundle. For based paths $\{\psi_t^{(1)}\}$ and $\{\psi_t^{(2)}\}$ with $\widetilde{\mathrm{Flux}}(\{\psi_t^{(1)}\}) = \widetilde{\mathrm{Flux}}(\{\psi_t^{(2)}\})$ we have a natural isomorphism*

$$HFN^*(\{\psi_t^{(1)}\}; L) \cong HFN^*(\{\psi_t^{(2)}\}; L).$$

Theorem 2.10. *If* $\widetilde{\mathrm{Flux}}(\{\psi_t\})$ *is sufficiently small we have*

$$HFN^*(\{\psi_t\}; L) \cong HN^{*+n}(\theta; L) \otimes_{\bar{\Lambda}_\theta} \Lambda_{\omega,\theta}$$

for any flat vector bundle L. *Here* $HN^*(\theta; L)$ *is the Novikov cohomology of* θ *with coefficients in* L.

Based on the above preparation, we give an outline of the proof of the flux conjecture. Let $U \subset H^1(M; \mathbb{R})$ be a neighborhood of the origin consisting of $\theta \in U$, which is represented by a sufficiently C^1-small closed 1-form such that Theorems 2.8 and 2.10 holds for the flux $[\theta]$. We may assume that U is symmetric with respect to the origin. It is enough to show the following.

Claim. $\Gamma_\omega \cap U = \{0\}$.

If it is false, there is a based loop $\{\psi_t\}$ in $\mathrm{Symp}_0(M, \omega)$ such that $\theta = \widetilde{\mathrm{Flux}}(\{\psi_t\})$ belongs to $(\Gamma_\omega \cap U) \setminus \{0\}$. Denote by $\{\psi_t^{-\theta}\}$ the based symplectic isotopy generated by the vector field $X_{-\theta}$ which is the symplectic dual of $-\theta$. Then $\{\psi'_t = \psi_t^{-\theta} \circ \psi_t\}$ is a based symplectic isotopy, the flux of which vanishes. Hence, we can deform $\{\psi'_t\}$ up to homotopy keeping end points fixed to a Hamiltonian path $\{\phi_t\}$. Thus we obtain a based Hamiltonian path $\{\phi_t\}$ and a based symplectic path $\{\psi_t^{-\theta}\}$ with $\psi_1^{-\theta} = \phi_1$. Since $\psi_t^\theta = (\psi_t^{-\theta})^{-1}$, $\Phi_t = \phi_t \circ \psi_t^\theta$ is a based loop in $\mathrm{Symp}_0(M, \omega)$, which induces an isomorphism $\Phi\colon \gamma(t) \in LM \mapsto \Phi_t(\gamma(t)) \in LM$. It is clear that Φ restricts to one-to-one correspondence between 1-periodic orbits of $\{\phi_t^{-\theta}\}$ and 1-periodic orbits of $\{\phi_t\}$. Note that the former are constant loops at zeros of θ, since we assumed that θ is sufficiently C^1-small. On the other hand, Theorem 2.2 guarantees the existence of contractible 1-periodic orbits of $\{\phi_t\}$ as we noted there. Hence, Φ preserves the component of LM consisting of contractible loops. We have the following:

Lemma 2.11. $\Phi^*\alpha_{\{\phi_t\}} = \alpha_{\{\psi_t^{-\theta}\}}$.

As a consequence, we find that $\Phi\colon LM \to LM$ admits a lift $\tilde{\Phi}\colon \tilde{L}^{-\theta}M \to \tilde{L}^0 M$. Note also that Φ_t preserves the homotopy class of almost complex structures compatible with ω, hence $c_1(M)(u) = c_1(M)[\Phi_\#(u)]$. Here $u\colon S^1 \times S^1 \to M$ and $\Phi_\#(u)(s,t) = \Phi_t(u(s,t))$. Therefore $\tilde{\Phi}$ induces an isomorphism between the Floer–Novikov cohomology of $\{\psi_t^{-\theta}\}$ and the Floer cohomology of $\{\phi_t\}$. (Φ also induces an isomorphism between the moduli spaces of gradient trajectories in the sense of Kuranishi structures, after choosing almost compatible structures appropriately.) We can also see that $\tilde{\Phi}$ induces an isomorphism between the Novikov rings $\Lambda_{\omega,-\theta}$ and $\Lambda_\omega = \Lambda_{\omega,0}$. Namely, we find

Proposition 2.12. *Let* $L \to M$ *be an arbitrary flat vector bundle. Then there exists* $c \in \mathbb{Z}$ *such that*

$$\tilde{\Phi}_*\colon HFN^*(\{\psi_t^{-\theta}\}; L) \cong HF^{*+c}(\{\phi_t\}; L).$$

Since $-\theta$ is sufficiently C^1-small, Theorem 2.10 implies that

$$HFN^*(\{\psi_t^{-\theta}\}; L) \cong HN^{*+n}(-\theta; L) \otimes_{\bar{\Lambda}_{-\theta}} \Lambda_{\omega,-\theta}.$$

On the other hand, we have

$$HF^*(\{\phi_t\}; L) \cong H^{*+n}(M; L) \otimes \Lambda_\omega.$$

Now we choose $L \to M$ as the flat real line bundle $L_{\varepsilon\theta}$ associated to $\ell \in \pi_1(M) \mapsto \exp(\varepsilon \int \ell^*\theta) \in \mathbb{R}^*$. Then the pull back $\pi^*L_{\varepsilon\theta}$ of $L_{\varepsilon\theta}$ by $\pi\colon \overline{M} \to M$, which is used to define the Novikov cohomology of $\pm\theta$, becomes trivial as a flat bundle. Hence $HN^*(-\theta; L_{\varepsilon\theta})$ is isomorphic to $HN^*(-\theta; \mathbb{R})$ after forgetting the module structure over the Novikov ring. On the other hand, for ordinary cohomology, we have the jumping phenomenon at $\varepsilon = 0$, i.e., since θ is not an exact 1-form, $H^0(M; L_{\varepsilon\theta}) = 0$ for $\varepsilon \neq 0$ while $H^0(M; \mathbb{R}) = \mathbb{R}$ for $\varepsilon = 0$. Based on this observation, we can derive a contradiction. Hence the flux conjecture is proved.

Theorem 2.13. *The flux conjecture holds for any closed symplectic manifolds.*

Remark 2.14. The action of Hamiltonian loops on Floer cohomologies was studied by Seidel [44]. Viterbo [47] developed the theory of generating functions and explored applications to symplectic invariants. Y. G. Oh is the first to apply the Floer theoretical framework to Hofer's geometry [34], [35], partly inspired by the work of Chekanov [3] to be mentioned later. Seidel's work also stimulated progress in the study of Hofer's geometry, e.g., Entov's work [7] and Schwarz [42]. Oh generalized Schwarz's result to closed symplectic manifolds which are not necessarily symplectically aspherical, cf. Oh's contribution to this proceedings. Based on this generalization, Entov and Polterovich constructed in [8] an $\mathbb{R}$-valued quasi-homomorphism from (the universal covering group of) $\mathrm{Ham}(M, \omega)$.

There are different kinds of development from those mentioned in this section. For example, Viterbo applied the Floer cohomology to a problem in real algebraic geometry and proved that hyperbolic manifolds cannot be realized as the real part of "sufficiently positively curved" complex projective manifolds; cf. [22].

3. Floer theory for Lagrangian submanifolds

3.1. Fundamental construction. Let L_0, L_1 be closed embedded Lagrangian submanifolds in a closed symplectic manifold (P, Ω). We assume that L_0 and L_1 intersect transversely. Consider the path space $\mathcal{P}(L_0, L_1) = \{\gamma\colon [0, 1] \to P \mid \gamma(0) \in L_0,\ \gamma(1) \in L_1\}$ and define the action 1-form $\alpha = \alpha_{L_0,L_1}$ by

$$\alpha_{L_0,L_1}(\xi) = \int_0^1 \Omega(\xi(t), \dot{\gamma}(t))\, dt \quad \text{for } \xi = \{\xi(t)\} \in T_\gamma \mathcal{P}(L_0, L_1).$$

Then α_{L_0,L_1} is a "closed 1-form". In fact, a local primitive function around γ_0 is given by

$$\mathcal{A}^{\mathrm{loc}}_{L_0,L_1}(\gamma) = \int_{[0,1]\times[0,1]} w^*\Omega,$$

where $w\colon [0,1]\times[0,1] \to P$ such that $w(s,i) \in L_i$ for $i = 0, 1$, $w(0,t) = \gamma_0(t)$ and $w(1,t) = \gamma(t)$. As long as the image of w is contained in a small neighborhood of γ_0, $\mathcal{A}^{\mathrm{loc}}_{L_0,L_1}$ is well defined.

Before going further, we clarify the relation to the case of symplectomorphisms. Let ϕ be a symplectomorphism of (M,ω). Then its graph Γ_ϕ is a Lagrangian submanifold in $(M\times M, -\omega\oplus\omega)$. Denote by Δ the diagonal subset, which is the graph of the identity. Then we have the following identification:

$$G\colon L^\phi M \to \mathcal{P}(\Gamma_\phi, \Delta);\ \sigma(t) \mapsto \left(\sigma\left(1-\frac{t}{2}\right), \sigma\left(\frac{t}{2}\right)\right),$$

which satisfies $G^*\alpha_{\Gamma_\phi,\Delta} = \alpha_\phi$. In this way, the construction in this section is a generalization of the one in the previous section.

Pick a compatible almost complex structure J to equip $\mathcal{P}(L_0, L_1)$ with L^2-metric. Then the locally gradient flow line for α_{L_0,L_1} is described by $u\colon \mathbb{R}\times[0,1] \to P$ with $u(\tau, i) \in L_i$ for $i = 0, 1$, which satisfies

$$\frac{\partial u}{\partial \tau} + J(u)\frac{\partial u}{\partial t} = 0.$$

Existence of the limits $\lim_{\tau\to\pm\infty} u(\tau,t) \in L_0\cap L_1$ is equivalent to the condition that the energy $E(u)$ is finite. Note also that the zeros of α_{L_0,L_1} are exactly the constant paths at $L_0\cap L_1$.

In [9]–[13], Floer realized the idea of constructing an analogue of Morse complex for the action functional under the assumption that $\pi_2(P, L_i) = 0$ and that L_1 is a Hamiltonian deformation of L_0. In this situation the action functional admits a primitive function on $\mathcal{P}(L_0, L_1)$ and the grading of $L_0\cap L_1$, called the Maslov–Viterbo index $\mu = \mu_{L_0,L_1}$, is well defined with values in $\mathbb{Z}$. Define $CF^*(L_1, L_0)$ by the $\mathbb{Z}/2\mathbb{Z}$-module freely generated by $L_0\cap L_1$. Counting gradient flow lines connecting critical points of $\mathcal{A}_{L_0,L_1}$, we define the coboundary operator $\delta\colon CF^*(L_1, L_0) \to CF^{*+1}(L_1, L_0)$ by

$$\delta\langle p\rangle = \sum \#\mathcal{M}(p,q)\langle q\rangle,$$

where q runs over $L_0\cap L_1$ such that $\mu(q) = \mu(p) + 1$, and $\mathcal{M}(p,q)$ is the moduli space of gradient flow lines, which we call connecting orbits, of $\mathcal{A}_{L_0,L_1}$ from p to q. Under the above assumption, for a generic choice of J, the moduli space $\mathcal{M}(p,q)$ is shown to be compact if $\mu(q) - \mu(p) = 1$. If $\mu(q) - \mu(p) = 2$, $\mathcal{M}(p,q)$ may not be compact, but its end is described as the union of $\mathcal{M}(p,r)\times\mathcal{M}(r,q)$ over $r \in L_0\cap L_1$

such that $\mu(r) - \mu(p) = 1$. Hence we find that $\delta \circ \delta = 0$ and obtain the Floer complex $(CF^*(L_1, L_0), \delta)$. We denote by $HF^*(L_1, L_0)$ the resulting cohomology. It is also shown that the Floer cohomology is invariant under Hamiltonian deformation of Lagrangian submanifolds. If L_1 is a sufficiently small Hamiltonian deformation of L_0, L_1 is regarded as the graph of a C^2-small Morse function on L_0 in T^*L_0. The Morse gradient trajectories appear as Floer connecting orbits. Although, in general, there may exist non-small Floer connecting orbits, the assumption that $\pi_2(P, L_i) = 0$ excludes such a possibility. Hence $HF^*(L_1, L_0)$ is isomorphic to $H^*(L_0; \mathbb{Z}/2\mathbb{Z})$ up to a shift in the grading. It is worth mentioning that Hofer [19], [20] developed an idea similar to Floer's and established the Lagrangian intersection property under the assumption that $\pi_2(P, L) = 0$.

Without the assumption that $\pi_2(P, L_i) = 0$, there arise some problems in the above argument. As we explain below, $\delta \circ \delta$ may not vanish[1], in general. It was Y. G. Oh [30], [31], [32] who extended Floer's construction to the case that L_i are monotone and their minimal Maslov number is at least 3. (He also computed Floer cohomology for some cases, e.g., $\mathbb{R}P^n \subset \mathbb{C}P^n$.) In general, the difficulties are caused by J-holomorphic discs with boundary on L_i as well as J-holomorphic spheres which arise as "bubbles" from sequences of connecting orbits with bounded energies. As in the case of symplectomorphisms, the bubbling-off of J-holomorphic spheres is expected to occur in real codimension 2 and does not cause any essential difficulty, which can be handled by Kuranishi structures. However, the bubbling-off of J-holomorphic discs occurs in real codimension 1 and we cannot avoid it, in general. If we restrict ourselves to some portion of $\mathcal{P}(L_0, L_1)$, on which the range of the action functional is sufficiently narrow, then there do not appear effects from J-holomorphic discs and J-holomorphic spheres. In fact, Chekanov [3] gave an alternative proof for the non-degeneracy of Hofer's distance on $\mathrm{Ham}(M, \omega)$ based on such an idea.

As we noticed, the bubbling-off of J-holomorphic discs is a codimension 1 phenomenon, hence we cannot, in general, avoid such a bubbling-off phenomenon from the moduli space $\mathcal{M}(p, q)$ even though $\mu(q) - \mu(p) \leq 2$. In order to understand how $\delta \circ \delta = 0$ fails to hold, we study all J-holomorphic discs systematically. From now on we follow our joint work with K. Fukaya, Y. G. Oh and H. Ohta [15]. Firstly, we arrange elements of $\pi_2(P, L)$[2], which are represented as the union of J-holomorphic discs $w\colon (D^2, \partial D^2) \to (P, L)$ and J-holomorphic spheres $v\colon S^2 \to P$ as $\beta_0 = 0, \beta_1, \beta_2, \ldots$ such that $\int_{\beta_i} \Omega \leq \int_{\beta_{i+1}} \Omega$ and $\int_{\beta_i} \Omega \to +\infty$ as $i \to +\infty$. This can be done with the so-called Gromov weak compactness. Denote by $\mu(w)$ the Maslov index of $(w^*TP, w|_{\partial D^2})^*TL) \to (D^2, \partial D^2)$. Denote by $\mathcal{M}_{k+1}(\beta)$ the moduli space of J-holomorphic discs[3] which represent class β, with $k+1$ marked points on ∂D^2. Then the moduli space $\mathcal{M}_{k+1}(\beta)$ is of dimension $n + \mu(\beta) + k - 2$, where

[1] I heard from A. Sergeev that Floer himself had (certainly) noticed this fact. This phenomenon is not only a bad news. We used this fact in [37].

[2] More precisely, we work with the image of $\pi_2(P, L)$ in $H^2(P, L; \mathbb{Z})$.

[3] More precisely, we use the stable maps from the prestable Riemann surface with 1 boundary component of genus 0.

$n = \dim L$. In general, the transversality, i.e., the surjectivity of the linearization of the J-holomorphic curve equation, may not hold. In order to overcome this trouble, we use the framework of Kuranishi structure. Since we use the multi-valued perturbation technique, we need a compatible system of orientations on various moduli spaces. However, the moduli space of J-holomorphic discs may not be orientable[4], in general. Therefore we assume the relative spin condition for Lagrangian submanifolds as follows. Pick a triangulation of L and extend it to a triangulation of P.

Definition 3.1 (Relative spin structure). Let L be a Lagrangian submanifold. If there is a cohomology class $w \in H^2(P; \mathbb{Z}/2\mathbb{Z})$ such that $w_2(L)$ is the restriction of w to L, we call L relatively spin. Under this condition, there is an orientable vector bundle V on the 3-skeleton $P^{(3)}$ of P with $w_2(V) = w$. A relative spin structure for L is the tuple of w, V and a spin structure on the restriction of $TL \oplus V$ to $L \cap P^{(2)}$. A relative spin structure on (L_0, L_1) is the above tuple which is chosen in common for L_i, $i = 0, 1$.

Then we have the following:

Theorem 3.2. (1) *A relative spin structure on L determines a canonical orientation on the moduli spaces $\mathcal{M}_{k+1}(\beta)$, which satisfies a certain compatibility condition under the gluing operation.*

(2) *A relative spin structure on (L_0, L_1) determines a canonical orientation on the moduli spaces $\mathcal{M}(p, q)$ of connecting orbits, which satisfies a certain compatibility condition under the gluing operation.*

From now on, we assume that a Lagrangian submanifold L or a pair (L_0, L_1) of Lagrangian submanifolds are equipped with a relative spin structure. We work with $\mathbb{Q}$-coefficients rather than $\mathbb{Z}/2\mathbb{Z}$-coefficients. Clearly, a spin structure on L gives a relative spin structure with a trivial bundle V.

We define obstruction classes for L to define Floer cohomology by inductive steps as follows[5]. Start with β_1, the first non-trivial case. Since the bubbling-off does not happen in $\mathcal{M}_1(\beta_1)$, the evaluation map $ev_0 \colon \mathcal{M}_1(\beta_1) \to L$ is a cycle with $\mathbb{Q}$-coefficients. This cycle represents the first obstruction class[6] $o_1 = o(\beta_1)$. Suppose that $o_i = o(\beta_i)$ is defined for $i = 1, \dots, k$ and there exist $\mathbb{Q}$-chains $\mathcal{B}_i$ in L such that $o_i = (-1)^n \partial \mathcal{B}_i$ for $i = 1, \dots, k$. (We call such a system of $\mathcal{B}_i, i = 1, 2, \dots$ a bounding chain.) We define the next obstruction class $o_{k+1} = o(\beta_{k+1})$ as follows. The moduli space $\mathcal{M}_{k+1}(\beta)$ may have codimension 1 boundary, hence may not be a cycle. So we try to glue other (moduli) spaces along boundaries so that we finally obtain a cycle. Consider the moduli space $\mathcal{M}_{\ell+1}(\beta; \mathcal{B}_{i_1}, \dots, \mathcal{B}_{i_\ell})$ consisting of J-holomorphic discs w representing the class β such that $\beta_{k+1} = \beta + \sum_{j=1}^{\ell} \beta_{i_j}$ and intersecting $\mathcal{B}_{i_1}, \dots, \mathcal{B}_{i_\ell}$ along ∂D^2. The moduli space $\mathcal{M}_{\ell+1}(\beta; \mathcal{B}_{i_1}, \dots, \mathcal{B}_{i_\ell})$ is

[4] Vin de Silva independently studied this problem in [5].

[5] The idea of this construction was inspired by Kontsevich around 1997.

[6] In the next subsection we adopt cohomological convention. Thus we take the Poincaré dual of o_k.

described as the fiber product of the spaces with Kuranishi structures:

$$\mathcal{M}_{\ell+1}(\beta)\ {}_{ev_1,\dots,ev_\ell}\times \prod_{j=1}^{\ell} \mathcal{B}_{i_j}.$$

We can assign to them an orientation so that the union $\widehat{\mathcal{M}}_1(\beta_{k+1})$ of $\mathcal{M}_1(\beta_{k+1})$ and all possible $\mathcal{M}_{\ell+1}(\beta; \mathcal{B}_{i_1}, \dots, \mathcal{B}_{i_\ell})$ becomes a $\mathbb{Q}$-virtual cycle. Note that we have the evaluation map $ev_0\colon \mathcal{M}_{\ell+1}(\beta; P_{i_1}, \dots, P_{i_\ell}) \to L$ at the remaining marked point after taking the fiber product. Then $ev_0\colon \widehat{\mathcal{M}}_1(\beta_{k+1}) \to L$ is a $\mathbb{Q}$-cycle of L, which represents the obstruction class $o_{k+1} = o(\beta_{k+1})$. Then we can find the following:

Theorem 3.3. *Suppose that a pair (L_0, L_1) of Lagrangian submanifolds is equipped with a relative spin structure. If all the obstruction classes for L_i, $i = 0, 1$ are defined and vanish, then we can revise the definition of Floer's coboundary operator to obtain the Floer complex $(CF^*(L_1, L_0), \delta)$. Moreover, the Floer cohomology $HF^*(L_1, L_0)$ is invariant under Hamiltonian deformation of L_i.*

Our construction depends on the choice of bounding chains for L_i, $i = 0, 1$. The invariance under Hamiltonian deformations also requires a subtle argument. Namely, we must describe the relation of bounding chains under Hamiltonian deformation. These points are clarified in terms of the filtered A_∞-algebras associated to L_i, which we discuss in the next subsection. We may weaken the assumption that the obstruction classes vanish. One of them is the deformation using $\mathbb{Q}$-cycles in P. It may also happen that the effects of J-holomorphic discs with boundary on L_i, $i = 0, 1$ cancel each other. When all non-vanishing obstruction classes for L_i are of top dimension, i.e., dim L, then they are multiples of the fundamental class of L. We call the coefficient of the fundamental cycle as the potential function of L_i. If the potential function of L_i, $i = 0, 1$, coincide, they cancel each other in the construction of the Floer complex, hence the Floer cohomology. This is an extension of Oh's discovery that the Floer complex can be constructed for monotone Lagrangian submanifolds with minimal Maslov numbers are at least 2. Although we can define the Floer complex, hence the Floer cohomology under the assumption that all obstruction classes vanish, it is very difficult to compute it in general.

However, when L_1 is a Hamiltonian deformation of L_0, we can construct a certain spectral sequence with E_2-term being the ordinary cohomology with coefficients in the Novikov ring, which converges to the Floer cohomology, see Theorem 3.10 below.

3.2. The filtered A_∞-algebras associated to Lagrangian submanifolds. Based on [15], we describe the framework of the Floer theory for Lagrangian submanifolds. We generalize the idea of the construction of obstruction classes, which we mentioned in the previous subsection, and construct the filtered A_∞-algebras associated to Lagrangian submanifolds. We also include some applications at the end of this subsection.

We introduce the universal Novikov ring which we use from now on. Let R be a commutative ring with the unit. In this note, we mostly use the case that $R = \mathbb{Q}$. Let T and e be formal generators of degree 0 and 2, respectively. Set

$$\Lambda_{\mathrm{nov}} = \Big\{ \sum_{i=0}^{\infty} a_i T^{\lambda_i} e^{\mu_i} \mid a_i \in R,\ \lambda_i \in \mathbb{R},\ \mu_i \in \mathbb{Z},\ \lim_{i\to\infty} \lambda_i = \infty \Big\}.$$

If R is a field, the degree 0-part of Λ_{nov} is also a field. We also set

$$\Lambda_{0,\mathrm{nov}} = \Big\{ \sum_{i=0}^{\infty} a_i T^{\lambda_i} e^{\mu_i} \in \Lambda_{\mathrm{nov}} \mid \lambda_i \geq 0 \Big\}.$$

These rings Λ_{nov} and $\Lambda_{0,\mathrm{nov}}$ are complete with respect to the decreasing filtration by λ for T^{λ}. The Novikov rings, we mentioned before, are subrings of Λ_{nov}.

Now we shall present a rough idea of the construction of the A_∞-operations. Let $(f_i\colon P_i \to L),\ i = 1, \ldots, k$, be chains in L. We often abbreviate them as P_i. Take the fiber product

$$\mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k) = \mathcal{M}_{k+1}(\beta)_{ev_1,\ldots,ev_k} \times_{f_1,\ldots,f_k} \prod_{j=1}^{k} P_j.$$

We can give an orientation to these spaces with Kuranishi structure in such a way that the following construction works. Define a chain $(\mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k), ev_0)$ in L by taking the remaining marked point, i.e., $ev_0\colon \mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k) \to L$. For $k \geq 2$, we set

$$\mathfrak{m}_{k,\beta}(P_1, \ldots, P_k) = (\mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k), ev_0).$$

In the other cases, we set

$$\begin{aligned} \mathfrak{m}_{1,0}(P) &= (-1)^n \partial P, \\ \mathfrak{m}_{1,\beta}(P) &= (\mathcal{M}_2(\beta; P), ev_0), \quad \text{when } \beta \neq 0 \\ \mathfrak{m}_{0,\beta}(1) &= (\mathcal{M}_1(\beta), ev_0), \qquad \text{when } \beta \neq 0. \end{aligned}$$

In the last line, 1 is the unit of $R \subset \Lambda_{\mathrm{nov}}$, which is regarded as an element in $B_0C(L; \Lambda_{0,\mathrm{nov}})[1]$ below. We also set $\mathfrak{m}_{0,0}(1) = 0$. If we study the structure of compactifications of the moduli spaces $\mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k)$ in a heuristic way, we expect to obtain certain algebraic relations among these operations, the so-called A_∞-relations. However, when we perform this construction in a rigorous way, we encounter several problems, e.g., transversality of the moduli spaces, transversality for taking the fiber product, etc. So we have to clarify which class of chains of L we deal with and how to take the (multi-valued) perturbation for achieving transversality, etc. Here, we give some flavor of the argument. For details see [15].

First of all, we forget the effect of non-trivial J-holomorphic discs and consider only the contribution from $\beta = 0$ (classical case). Naively, $\mathfrak{m}_{2,0}(P_1, P_2)$ should be $P_1 \cap P_2$ up to sign. However, when $P_1 = P_2$, the transversality does not hold. Thus we are forced to perturb $\mathcal{M}_{3,0}(0; P_1, P_2)$ to define $\mathfrak{m}_{2,0}(P_1, P_2)$. (It is also necessary to take a suitable countable family of chains which spans a subcomplex of the chain complex of L. We also assume that its cohomology is isomorphic to the ordinary cohomology of L.) It causes a discrepancy between $\mathfrak{m}_{2,0}(\mathfrak{m}_{2,0}(P_1, P_2), P_3)$ and $\mathfrak{m}_{2,0}(P_1, \mathfrak{m}_{2,0}(P_2, P_3))$. Namely, $\mathfrak{m}_{2,0}$ does not satisfy the associativity. Nevertheless, the above discrepancy is described using $\mathfrak{m}_{3,0}(P_1, P_2, P_3)$, which is defined by the perturbation of $\mathcal{M}_{4,0}(0; P_1, P_2, P_3)$, as follows.

$$\begin{aligned}
&\mathfrak{m}_{2,0}(\mathfrak{m}_{2,0}(P_1, P_2), P_3) - (-1)^{\deg P_1}\mathfrak{m}_{2,0}(P_1, \mathfrak{m}_{2,0}(P_2, P_3)) \\
&\quad = -\{\mathfrak{m}_{1,0} \circ \mathfrak{m}_{3,0}(P_1, P_2, P_3) + \mathfrak{m}_{3,0}(\mathfrak{m}_{1,0}(P_1), P_2, P_3) \\
&\qquad - (-1)^{\deg P_1}\mathfrak{m}_{3,0}(P_1, \mathfrak{m}_{1,0}(P_2), P_3) \\
&\qquad + (-1)^{\deg P_1 + \deg P_2}(\mathfrak{m}_{3,0}(P_1, P_2, \mathfrak{m}_{1,0}(P_3))\}.
\end{aligned}$$

Here we define the degree of P by $\deg P = n - \dim P$ and work with the cohomological framework rather than the homological framework from now on. A series of similar formulae successively hold in higher order. We call these relations the A_∞-relations. We can show that this algebraic gadget, the A_∞-algebra, obtained by the chain level intersection theory is "equivalent" to the de Rham homotopy theory in the realm of A_∞-algebras.

Next we include the effect from non-trivial J-holomorphic discs. Then we first take a suitable countably generated subcomplex $C^*(L)$ of the (co)chain complex[7] and (multi-valued) perturbations of the moduli spaces $\mathcal{M}_{k+1}(\beta; P_1, \ldots, P_k)$ to define $\mathfrak{m}_{k,\beta}(P_1, \ldots, P_k)$. We assign the degree to $P \otimes T^\lambda e^\mu \in C^*(L; \Lambda_{\mathrm{nov}})$ by $\deg P + 2\mu$. We shift the degree as $C(L; \Lambda_{\mathrm{nov}})[1]^* = C^{*+1}(L; \Lambda_{\mathrm{nov}})$. Then we can easily see that

$$\mathfrak{m}_{k,\beta} \otimes T^{\int_\beta \Omega} e^{\mu(\beta)/2} \colon \bigotimes^k C(L; \Lambda_{\mathrm{nov}})[1]^* \to C(L; \Lambda_{\mathrm{nov}})[1]^*$$

shifts the degree by $+1$, in other words, they are operations of degree $+1$. Write

$$\mathfrak{m}_k = \sum_\beta \mathfrak{m}_{k,\beta} \otimes T^{\int_\beta} e^{\mu(\beta)/2}.$$

Write

$$BC[1]^* = \bigoplus_{k=0}^\infty B_k C[1]^*$$

and

$$B_k C[1]^* = B_k C(L; \Lambda_{0,\mathrm{nov}})[1]^* = \bigotimes^k C(L; \Lambda_{0,\mathrm{nov}})[1]^*,$$

[7] More precisely, we consider the quotient complex by identifying chains, which give the same current.

the bar construction of $C^* = C^*(L; \Lambda_{0,\mathrm{nov}})$. It is a free tensor coalgebra generated by the graded free module $C[1]^*$. We extend $\mathfrak{m}_k$ to $\widehat{\mathfrak{m}}_k \colon BC[1]^* \to BC[1]^*$ as a coderivation and define $\hat{d} = \sum \widehat{\mathfrak{m}}_k$. Then we find the following:

Theorem 3.4. $\hat{d} \circ \hat{d} = 0$.

The filtered A_∞-relations are the formulae which express the above equality in terms of $\mathfrak{m}_k$. We call $(BC(L; \Lambda_{0,\mathrm{nov}})[1], \hat{d})$ the filtered A_∞-algebra associated to the Lagrangian submanifold L. So far, this object depends on the choice of the compatible almost complex structure, the countably generated subcomplex $C(L)$, various (multi-valued) perturbations, etc. We can define the notion of (gapped filtered) A_∞-algebra morphisms, homotopy equivalences, homotopy units, etc., and find the following:

Theorem 3.5. (1) *The homotopy type of the filtered A_∞-algebra*

$$(BC(L; \Lambda_{0,\mathrm{nov}})[1], \hat{d})$$

depends only on the embedding of the Lagrangian submanifold $L \subset (P, \Omega)$. The fundamental cycle of L is a homotopy unit.

(2) *A symplectomorphism ψ of (P, Ω) induces a homotopy equivalence $\hat{\psi}$ between the filtered A_∞-algebras associated to L and $\psi(L)$.*

In fact, by the algebraic theory of the (filtered) A_∞-algebras, we can derive the A_∞-algebra structure, resp. the filtered A_∞-algebra structure on $H^*(L)$, resp. $H^*(L; \Lambda_{0,\mathrm{nov}})$. One of the advantages to work in the framework of (filtered) A_∞-algebras is that quasi-isomorphisms have homotopy inverses[8]. This is not true in the category of differential graded algebras.

In general, $\mathfrak{m}_0(1)$ may not vanish. From the A_∞-relation we have

$$\mathfrak{m}_1 \circ \mathfrak{m}_1(P) = -(\mathfrak{m}_2(\mathfrak{m}_0(1), P) + (-1)^{\deg P + 1}\mathfrak{m}_2(P, \mathfrak{m}_0(1))),$$

which means that $\mathfrak{m}_1 \circ \mathfrak{m}_1$ does not necessarily vanish. This is the obstruction to define the Floer cohomology, which we discussed in the previous subsection.

Let $b \in C(L; \Lambda_{0,\mathrm{nov}})[1]^0$ with positive energy, i.e, b contains only terms with T^λ with $\lambda > 0$ and set

$$\mathfrak{m}_k^b(P_1, \ldots, P_k) = \sum \mathfrak{m}_{k+\ell}(b, \ldots, b, P_1, b, \ldots, b, P_i, b, \ldots, b, P_k, b, \ldots, b),$$

where ℓ is the number of b's appearing above in all possible ways and the sum is taken over all possibilities. We define $\hat{d}^b$ using $\mathfrak{m}_k^b$ instead of $\mathfrak{m}_k$. Then $\hat{d}^b$ also satisfies the A_∞-relation $\hat{d}^b \circ \hat{d}^b = 0$. Write $e^b = 1 + b + b \otimes b + b \otimes b \otimes b + \cdots$. Then we find the following:

Theorem 3.6. *If there exists $b \in C(L; \Lambda_{0,\mathrm{nov}})[1]^0$ which satisfies $\hat{d}(e^b) = 0$, then we have $\mathfrak{m}_0^b(1) = 0$, hence $\mathfrak{m}_1^b \circ \mathfrak{m}_1^b = 0$.*

[8] We do not claim any priority in the unfiltered case.

We call the equation $\hat{d}(e^b) = 0$ the Maurer–Cartan equation for the filtered A_∞-algebra. If there is a solution b for the Maurer–Cartan equation, the complex $(C(L;\Lambda_{0,\mathrm{nov}}), \mathfrak{m}_1^b)$ and its extension $(C(L;\Lambda_{\mathrm{nov}}), \mathfrak{m}_1^b)$ are the Bott–Morse Floer complex in the case that $L = L_0 = L_1$. We denote the resulting cohomology groups by $HF^*((L,b);\Lambda_{0,\mathrm{nov}})$ and $HF^*((L,b);\Lambda_{\mathrm{nov}})$, respectively. For a bounding chain $\mathcal{B}_i$ in the previous subsection we set

$$b = \sum \mathcal{B}_i \otimes T^{\int_\beta \Omega} e^{\mu(\beta)/2}.$$

Then b is a solution of the Maurer–Cartan equation. There is a notion of the gauge equivalence relation among solutions of the Maurer–Cartan equation. We can see that the Floer cohomologies are isomorphic for gauge equivalent b and b'. Note that the filtered A_∞-morphism maps a solution of the Maurer–Cartan equation for the source to a solution of the Maurer–Cartan equation for the target. Hence, for a symplectomorphism ψ of (P,Ω), $\psi_*(b)$, the B_1-component of $\widehat{\psi}(e^b)$ is a solution of the Maurer–Cartan equation for $\psi(L)$ if b is a solution for L. With respect to them we have the following:

$$\widehat{\psi}:\ HF^*((L,b);\Lambda_{0,\mathrm{nov}}) \cong HF^*((\psi(L),\psi_*(b));\Lambda_{0,\mathrm{nov}}).$$

Now we consider a pair (L_0, L_1) of Lagrangian submanifolds. By counting Floer connecting orbits intersecting k chains in L_1 and ℓ chains in L_0, we define the operation

$$\begin{aligned}\mathfrak{n}_{k,\ell}:\ B_kC(L_1;\Lambda_{0,\mathrm{nov}})[1] \otimes C(L_1,L_0;\Lambda_{0,\mathrm{nov}}) \otimes B_\ell C(L_0;\Lambda_{0,\mathrm{nov}})[1]\\ \longrightarrow C(L_1,L_0;\Lambda_{0,\mathrm{nov}}).\end{aligned}$$

Using the filtered A_∞-algebra structures on L_1 and L_0 as well as $\mathfrak{n}_{k,\ell}$, we obtain the coderivation $\hat{d}_{(L_1,L_0)}$ on

$$BC(L_1;\Lambda_{0,\mathrm{nov}})[1] \otimes C(L_1,L_0;\Lambda_{0,\mathrm{nov}}) \otimes BC(L_0;\Lambda_{0,\mathrm{nov}})[1].$$

We have $\hat{d}_{(L_1,L_0)} \circ \hat{d}_{(L_1,L_0)} = 0$. We call $(BC(L_1;\Lambda_{0,\mathrm{nov}})[1] \otimes C(L_1,L_0;\Lambda_{0,\mathrm{nov}}) \otimes BC(L_0;\Lambda_{0,\mathrm{nov}})[1], \hat{d}_{(L_1,L_0)})$ the filtered A_∞-bimodule associated to the pair (L_0, L_1). More precisely, we say that it is a left $C(L_1;\Lambda_{0,\mathrm{nov}})$, right $C(L_0;\Lambda_{0,\mathrm{nov}})$ filtered A_∞-bimodule. Similar to the case of the filtered A_∞-algebras, we obtain the following:

Theorem 3.7. (1) *For a pair (L_0, L_1) of Lagrangian submanifold equipped with a relative spin structure as a pair, the filtered A_∞-bimodule is uniquely defined up to homotopy equivalences.*

(2) *A pair of Hamiltonian diffeomorphisms ϕ_i, $i = 0, 1$, induces a homotopy equivalence between the filtered A_∞-bimodules with coefficients in Λ_{nov} of (L_0, L_1) and $(\phi_0(L_0), \phi_1(L_1))$.*

If there exists solutions b_i of the Maurer–Cartan equations for L_i, we can revise the Floer coboundary operator as follows:

$$\delta^{b_1,b_0}(p) = \sum \mathfrak{n}_{k,\ell}(b_1, \dots b_1, p, b_0, \dots, b_0).$$

Then we have the following:

Theorem 3.8. *Let b_i be solutions of the Maurer–Cartan equations for L_i, $i = 0, 1$. Then $\delta^{b_1,b_0} \circ \delta^{b_1,b_0} = 0$ holds.*

We denote the resulting cohomology by $HF^*((L_1, b_1), (L_0, b_0); \Lambda_{0,\mathrm{nov}})$ and its coefficient extension to Λ_{nov} by $HF^*((L_1, b_1), (L_0, b_0); \Lambda_{\mathrm{nov}})$ Then we have the following:

Corollary 3.9. *Let b_i be solutions of the Maurer–Cartan equation for L_i and ϕ_i Hamiltonian diffeomorphisms of $(P.\Omega)$. Then (ϕ_1, ϕ_0) induces an isomorphism*

$$HF^*((L_1.b_1), (L_0, b_0); \Lambda_{\mathrm{nov}}) \cong HF^*((\phi_1(L_1), \phi_{1*}(b_1)), (\phi_0(L_0), \phi_{0*}(b_0)); \Lambda_{\mathrm{nov}}).$$

In a similar way to the case of filtered A_∞-algebras, if b_i is gauge equivalent to b_i', $i = 0, 1$, the corresponding Floer cohomologies are isomorphic. Suppose that $\mathfrak{m}_0(1) = c[L]$ for some $c \in \Lambda_{0,\mathrm{nov}}$. We set $c = \mathfrak{PO}(L)$, the potential function. If $\mathfrak{PO}(L_0) = \mathfrak{PO}(L_1)$ we can modify the above construction to obtain the Floer complex. For example, if L_0 is a Lagrangian submanifold such that $\mathfrak{m}_0(1) = c[L_0]$ and L_1 is a Hamiltonian deformation of L_0, then we can obtain the Floer cohomology for (L_0, L_1).

It is not easy to compute the Floer cohomology $HF^*((L_1, b_1), (L_0, b_0))$, even when $L_1 = \phi(L_0)$ and $b_1 = \phi_*(b_0)$ for some Hamiltonian diffeomorphism ϕ. In such a case we find that it is isomorphic to the Bott–Morse Floer cohomology $HF^*((L_0, b_0); \Lambda_{\mathrm{nov}})$. Using the energy filtration, we have a spectral sequence as follows.

Theorem 3.10. *There is a spectral sequence with E_2-term being $H^*(L; \Lambda_{0,\mathrm{nov}})$ and converging to $HF^*((L_0, b_0); \Lambda_{0,\mathrm{nov}})$.*

We can also use a cycle in the ambient space P to deform the filtered A_∞-algebra associated to L. Pick a cycle $\boldsymbol{b}$ in P. Consider the moduli space of stable maps with one boundary component. In addition to the $k + 1$ boundary marked points put ℓ interior marked points. Take the fiber product

$$\mathcal{M}_{k+1,\ell}(\beta; P_1, \dots, P_k) = \mathcal{M}_{k+1,\ell}(\beta) \times_{\prod^k L \times \prod^\ell P} \Big(\prod_{i=1}^k P_i\Big) \times \Big(\prod^\ell \boldsymbol{b}\Big).$$

Summing up these moduli spaces for all ℓ, we obtain the deformed operation $\mathfrak{m}_k^{\boldsymbol{b}}$. The corresponding $\hat{d}^{\boldsymbol{b}}$ gives a deformation of the filtered A_∞-algebra structure. We can also discuss the Maurer–Cartan equation for the deformed structure, gauge equivalences, etc. Thanks to this larger class of deformations, we have the following:

Theorem 3.11. *Let L be a relatively spin Lagrangian submanifold. If the embedding $L \subset P$ induces an injection $H^*(L;\mathbb{Q}) \to H^*(P;\mathbb{Q})$, there is a $\Lambda_{0,\mathrm{nov}}^+$-cycle*[9] *$\boldsymbol{b}$ of P such that the deformed Maurer–Cartan solution $\hat{d}^{\boldsymbol{b}}(e^b) = 0$ has a solution.*

The following theorem is a direct consequence.

Theorem 3.12. *Let L be a relatively spin Lagrangian submanifold. Suppose that the embedding $L \subset P$ induces an injection on homology with rational coefficients. Then, for any Hamiltonian diffeomorphism ϕ of (P, Ω), we have*

$$\#L \cap \phi(L) \geq \sum_p \operatorname{rank} H^p(L;\mathbb{Q}).$$

Note that the graph of a Hamiltonian diffeomorphism satisfies the above assumption, hence Theorem 3.12 is a generalization of Theorem 2.2. Although the complete computation is difficult, there are cases where we have the non-vanishing result.

Theorem 3.13. *Let L be a relatively spin Lagrangian submanifold. Suppose that there is a $\Lambda_{0,\mathrm{nov}}^+$-cycle $\boldsymbol{b}$ in P and $b \in C(L;\Lambda_{0,\mathrm{nov}})[1]^0$ such that $\hat{d}^{\boldsymbol{b}}(e^b) = 0$. Suppose also that the Maslov index of any J-holomorphic disc with boundary on L is non-positive. Then, after adding correction terms which are of positive energy, the cycle $[pt]$ and the cyle $[L]$ become linearly independent, non-trivial cohomology classes in $HF^*((L,\mathfrak{b});\Lambda_{\mathrm{nov}})$.*

Here we denote by $\mathfrak{b}$ the pair $(\boldsymbol{b}, b)$. When the Maslov class μ vanishes for L, all obstruction classes belong to $H^2(L;\mathbb{Q})$. Hence we obtain the following:

Theorem 3.14. *Let L be a relatively spin Lagrangian submanifold with vanishing Maslov class such that $H^2(L;\mathbb{Q}) = 0$. Then, for any Hamiltonian diffeomorphism ϕ, $L \cap \phi(L) \neq \emptyset$. Moreover, there is $p \in L \cap \phi(L)$ with Viterbo–Maslov index* 0.

Thomas and Yau [46] used this theorem to establish the uniqueness of special Lagrangian homology spheres. From an opposite viewpoint, if L is a relatively spin Lagrangian submanifold with vanishing second rational cohomology and admits a Hamiltonian diffeomorphism ϕ such that $L \cap \phi(L) = 0$, then the Maslov class μ_L does not vanish. For instance, we have the following:

Theorem 3.15. *Let L be a Lagrangian submanifold in the symplectic vector space $(\mathbb{R}^{2n}, \omega_{can})$. If $H^2(L;\mathbb{Q}) = 0$ then $\mu_L \neq 0$. Moreover, its minimal Maslov number is at most $n+1$.*

Some results in a similar spirit were also obtained by Biran and Cieliebak [2]. Y. G. Oh obtained a more precise upper bound for the minimal Maslov number for Lagrangian tori up to a certain dimension [33]. Once we know that there exists a Hamiltonian diffeomorphism ϕ of (P, Ω) such that $L \cap \phi(L) = \emptyset$, either some

[9] $\Lambda_{0,\mathrm{nov}}^+ = \{\sum a_i T^{\lambda_i} e^{\mu_i} \in \Lambda_{\mathrm{nov}} | \lambda_i > 0\}$.

obstruction class does not vanish, or some differential in the spectral sequence in Theorem 3.10 is non-trivial. In each case we obtain the existence of non-trivial J-holomorphic discs with boundary on L. Thus, for example, we can find that any Lagrangian submanifolds in symplectic vector spaces are not exact. Namely, the restriction of the Liouville form $\lambda = \sum p_i dq^i$ to L is not an exact 1-form on L (Gromov).

Finally we discuss an analogue of the flux conjecture for Lagrangian submanifolds. Denote by $\mathrm{Lag}(L)$ the space of all Lagrangian submanifolds which are Lagrangian isotopic to L with C^1-topology. Consider the quotient $\mathrm{Lag}(L)/\mathrm{Ham}(P, \Omega)$ by the obvious action of $\mathrm{Ham}(P, \Omega)$. The question is whether $\mathrm{Lag}(L)/\mathrm{Ham}(P, \Omega)$ is Hausdorff or not. This is false in general. In fact, Chekanov's example in [4] provides a counterexample. In his example the Maslov class is non-zero. As an application of our theory [15] we find the following result which is an analogue to Theorem 2.4 (the case that the Chern number is 0).

Theorem 3.16. *Let L be a relatively spin Lagrangian submanifold L with vanishing Maslov class. Suppose that the (deformed) Maurer–Cartan equation for L has a solution. If $L' = \phi(L)$, for some $\phi \in \mathrm{Ham}(P, \Omega)$, is sufficiently C^1-close to L, then L' is regarded as the graph of an exact 1-form on L via Weinstein's tubular neighborhood theorem.*

We expect that $\mathrm{Lag}(L)/\mathrm{Ham}(P, \Omega)$ is Hausdorff under the above assumption.

Finally, we make a remark that if L is a so-called semi-positive Lagrangian submanifold, we can work with $\mathbb{Z}/2\mathbb{Z}$-coefficients rather than $\mathbb{Q}$-coefficients. We do not need the relative spin condition in this case. There is also an approach to the Floer cohomology with $\mathbb{Z}$-coefficients [17]. There are also applications in relation to "mirror symmetry" which we do not discuss here.

Acknowledgement. I would like to thank my collaborators, K. Fukaya, Y. G. Oh, H. Ohta in [15] and H. V. Le in [27].

References

[1] Banyaga, A., Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique. *Comment. Math. Helv.* **53** (1978), 174–227.

[2] Biran, P., and Cieliebak, K., Lagrangian embeddings into subcritical Stein manifolds. *Israel J. Math.* **127** (2002), 221–244.

[3] Chekanov, Y., Lagrangian intersections, symplectic energy and areas of holomorphic curves. *Duke Math. J.* **95** (1998), 213–226.

[4] Chekanov, Y., Lagrangian torus in a symplectic vector space and global symplectomorphisms. *Math. Z.* **223** (1996), 547–559.

[5] de Silva, V., Products in the symplectic Floer homology of Lagrangian intersections. Ph.D. thesis, Oxford University, 1998.

[6] Dostoglou, S., and Salamon, D., Self-dual instantons and holomorphic curves. *Ann. of Math.* (2) **139** (1994), 581–640.

[7] Entov, M., K-area, Hofer metric and geometry of conjugacy classes in Lie groups. *Invent. Math.* **146** (2001), 93–141.

[8] Entov, M., and Polterovich, L., Calabi quasimorphisms and quantum homology. *Internat. Math. Res. Notices* **30** (2003), 1635–1676.

[9] Floer, A., Morse theory for lagrangian intersections. *J. Differential Geom.* **28** (1988), 513–547.

[10] Floer, A., The unregularized gradient flow of the symplectic action. *Comm. Pure Appl. Math.* **41** (1988), 775–813.

[11] Floer, A., A relative Morse index for the symplectic action. *Comm. Pure Appl. Math.* **41** (1988), 393–407.

[12] Floer, A., Witten's complex and infinite dimensional Morse theory. *J. Differential Geom.* **30** (1989), 207–221.

[13] Floer, A., Cup length estimate on lagrangian intersections. *Comm. Pure Appl. Math.* **42** (1989), 335–357.

[14] A. Floer, Holomorphic spheres and symplectic fixed points. *Comm. Math. Phys.* **120** (1989), 575–611.

[15] Fukaya, K., Oh, Y. G., Ohta, H., and Ono, K., Lagrangian intersection Floer theory – obstruction and anomaly. Preprint 2000 and revised version, in preparation.

[16] Fukaya, K., and Ono, K., Arnold conjecture and Gromov-Witten invariant. *Topology* **38** (1999), 933–1048.

[17] Fukaya, K., and Ono, K., Floer homology and Gromov-Witten invariant over integer for general symplectic manifolds. In *Taniguchi Conference on Mathematics Nara '98* (ed. by Masaki Maruyama and Toshikazu Sunada), Adv. Stud. Pure Math. 31, Mathematical Society of Japan, Tokyo 2001, 75–91.

[18] Gromov, M., Pseudoholomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), 307–347.

[19] Hofer, H., Lagrangian embeddings and critical point theory. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **2** (1985), 407–462.

[20] Hofer, H., Lusternik-Schnirelmann theory for Lagrangian intersections. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **5** (1988), 465–499.

[21] Hofer, H., and Salamon, D., Floer homology and Novikov rings. In *The Floer memorial volume* (ed. by H. Hofer, C. Taubes, A. Weinstein and E. Zehnder), Progr. Math. 133, Birkhäuser, Basel 1995, 483–524.

[22] Kharlamov, V., Variétés de Fano réels (d'après C. Viterbo). Séminaire Bourbaki, 1999/2000, *Astérisque* **276** (2002), 189–206.

[23] Kontsevich, M., Enumeration of rational curves by torus action. In *Moduli space of curves* (ed. by H. Dijkgraaf, C. Faber, G. v. d. Geer), Progr. Math. 129, Birkhäuser, Basel 1995, 335–368.

[24] Kontsevich, M., and Manin, Y., Gromov-Witten classes, quantum cohomology and enumerative geometry. *Comm. Math. Phys.* **164** (1994), 525–562.

[25] Lalonde, F., , McDuff, D., and Polterovich, L., On the flux conjectures. In *Geometry, topology and dynamics*, CRM Proc. Lecture Notes 15, Amer. Math. Soc., Providence, RI, 1998, 69–85.

[26] Lalonde, F., , McDuff, D., and Polterovich, L., Topological rigidity of Hamiltonian loops and quantum homology. *Invent. Math.* **135** (1999), 369–385.

[27] Lê, H. V., and Ono, K., Symplectic fixed points, the Calabi invariant and Novikov homology. *Topology* **34** (1995), 155–176.

[28] Li, J., and Tian, G., Virtual moduli cycles and Gromov-Witten invariants of general symplectic manifolds. In *Topics in symplectic* 4*-manifolds*, First Int. Press Lect. Ser. 1, International Press, Cambridge, MA, 1998, 47–83.

[29] Liu, G., and Tian, G., Floer homology and Arnold conjecture. *J. Differential Geom.* **49** (1998), 1–74.

[30] Oh, Y. G., Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, I. *Comm. Pure Appl. Math.* **46** (1993), 949–994 Addendum ibid. 48 (1995), 1299–1302.

[31] Oh, Y. G., Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, II. *Comm. Pure Appl. Math.* **46** (1993), 995–1012

[32] Oh, Y. G., Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, III. In *The Floer memorial volume* (ed. by H. Hofer, C. Taubes, A. Weinstein and E. Zehnder), Progr. Math. 133, Birkhäuser, Basel 1995, 555–573.

[33] Oh, Y. G., Floer cohomology, spectral sequences and the Maslov class of Lagrangian embeddings. *Internat. Math. Res. Notices* **7** (1996), 305–346.

[34] Oh, Y. G., Symplectic topology as the geometry of the action functional, I. *J. Differential Geom.* **46** (1997), 499–577.

[35] Oh, Y. G., Symplectic topology as the geometry of the action functional, II. *Commun. Anal. Geom.* **7** (1999), 1–55.

[36] Ono, K., On the Arnold conjecture for weakly monotone symplectic manifolds. *Invent. Math.* **119** (1995), 519–537.

[37] Ono, K., Lagrangian intersection under legendrian deformations. *Duke Math. J.* **85** (1996), 209–225.

[38] Ono, K., Floer-Novikov cohomology and the flux conjecture. Preprint.

[39] Ono, K., Floer-Novikov cohomology and symplectic fixed points. To appear in *Proceedings of Workshop on Symplectic Topology, Stare Jablonki, 2004*.

[40] Polterovich, L., Monotone Lagrangian submanifolds of linear spaces and the Maslov class in cotangent bundles. *Math. Z.* **207** (1991), 217–222.

[41] Ruan, Y., Virtual neighborhoods and pseudoholomorphic curves. *Turkish J. Math.* **23** (1999), 161–231.

[42] Schwarz, M., On the action spectrum for closed symplectically aspherical manifolds. *Pacific J. Math.* **193** (2000), 419–461.

[43] Seidel, P., Symplectic Floer homology and the mapping class group. *Pacific J. Math.* **206** (2002), 219–229.

[44] Seidel, P., π_1 of symplectic automorphism groups and invertibles in quantum cohomology rings. *Geom. Funct. Anal.* **7** (1997), 1046–1095.

[45] Siebert, B., Gromov-Witten invariants for general symplectic manifolds. Preprint, 1996.

[46] Thomas, R., and Yau, S. T., Special Lagrangians, stable bundles and mean curvature flow. *Comm. Geom. Anal.* **10** (2002), 1075–1113.

[47] Viterbo, C., Symplectic topology as the geometry of generating functions. *Math. Ann.* **292** (1992), 685–710.

Department of Mathematics, Hokkaido University, Sapporo, 060-0810, Japan
E-mail: ono@math.sci.hokudai.ac.jp

Heegaard diagrams and Floer homology

Peter Ozsváth* and Zoltán Szabó†

Abstract. We review the construction of Heegaard–Floer homology for closed three-manifolds and also for knots and links in the three-sphere. We also discuss three applications of this invariant to knot theory: studying the Thurston norm of a link complement, the slice genus of a knot, and the unknotting number of a knot. We emphasize the application to the Thurston norm, and illustrate the theory in the case of the Conway link.

Mathematics Subject Classification (2000). 53D, 57R.

Keywords. Heegaard diagrams, Floer homology, Thurston norm.

1. Heegaard–Floer homology of three-manifolds

Floer homology was initially introduced by Floer to study questions in Hamiltonian dynamics [8]. The basic set-up for his theory involves a symplectic manifold (M, ω), and a pair of Lagrangian submanifolds L_0 and L_1. His invariant, *Lagrangian Floer homology*, is the homology group of a chain complex generated freely by intersection points between L_0 and L_1, endowed with a differential which counts pseudo-holomorphic disks. This chain complex arises from a suitable interpretation of the Morse complex in a certain infinite-dimensional setting.

Soon after formulating Lagrangian Floer homology, Floer realized that his basic principles could also be used to construct a three-manifold invariant, *instanton Floer homology*, closely related to Donaldson's invariants for four-manifolds. In this version, the basic set-up involves a closed, oriented three-manifold Y (satisfying suitable other topological restrictions on Y; for example, the theory is defined when Y has trivial integral first homology). Again, one forms a chain complex, but this time the generators are $\mathrm{SU}(2)$ representations of the fundamental group of Y (or some suitable perturbation thereof), and the differentials count anti-self-dual Yang–Mills connections on the product of Y with the real line. This invariant plays a crucial role in Donaldson's invariants for smooth four-manifolds: for a four-manifold-with-boundary, the relative Donaldson invariant is a homology class in the instanton Floer homology groups of its boundary [4].

In the present note, we will outline an adaptation of Lagrangian Floer homology, *Heegaard–Floer homology*, which gives rise to a closed three-manifold invariant [33], [32]. This invariant also fits into a four-dimensional framework [27]. There is a

*PSO was supported by NSF grant number DMS-0505811.

†ZSz was supported by NSF grant number DMS-0406155.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

related invariant of smooth four-manifolds, and indeed relative invariants for this four-manifold invariant take values in the Heegaard–Floer homology groups of its boundary.

A Heegaard diagram is a triple consisting of a closed, oriented two-manifold Σ of genus g, and a pair of g-tuples of embedded, disjoint, homologically linearly independent curves $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_g\}$ and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_g\}$. A Heegaard diagram uniquely specifies a three-manifold, obtained as a union of two genus g handlebodies U_α and U_β. In U_α, the curves α_i bound disks, while in U_β, the curves β_i bound disks. We associate to this data a suitable version of Lagrangian Floer homology.

Our ambient manifold in this case is the g-fold symmetric product of Σ, the set of unordered g-tuples of points in Σ. This space inherits a natural complex structure from a complex structure over Σ. Inside this manifold, there is a pair of g-dimensional real tori, $\mathbb{T}_\alpha = \alpha_1 \times \cdots \times \alpha_g$ and $\mathbb{T}_\beta = \beta_1 \times \cdots \times \beta_g$. We fix also a reference point

$$w \in \Sigma - \alpha_1 - \cdots - \alpha_g - \beta_1 - \cdots - \beta_g.$$

This gives rise to a subvariety $V_w = \{w\} \times \mathrm{Sym}^{g-1}(\Sigma) \subset \mathrm{Sym}^g(\Sigma)$. We consider the chain complex generated by intersection points $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$. Concretely, an intersection point of $\mathbb{T}_\alpha$ and $\mathbb{T}_\beta$ corresponds to a permutation σ in the symmetric group on g letters, together with a g-tuple of points $\boldsymbol{x} = (x_1, \dots, x_g)$ with $x_i \in \alpha_i \cap \beta_{\sigma(i)}$.

The differential again counts holomorphic disks; but some aspect of the homotopy class of the disk is recorded. We make this precise presently. For fixed $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, let $\pi_2(\boldsymbol{x}, \boldsymbol{y})$ denote the space of homotopy classes of Whitney disks connecting $\boldsymbol{x}$ to $\boldsymbol{y}$, i.e. continuous maps of the unit disk $\mathbb{D} \subset \mathbb{C}$ into $\mathrm{Sym}^g(\Sigma)$, mapping the part of the boundary of $\mathbb{D}$ with negative resp. positive real part to $\mathbb{T}_\alpha$ resp. $\mathbb{T}_\beta$, and mapping i resp. $-i$ to $\boldsymbol{x}$ resp. $\boldsymbol{y}$. The algebraic intersection number of $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ with the subvariety V_w determines a well-defined map

$$n_w \colon \pi_2(\boldsymbol{x}, \boldsymbol{y}) \longrightarrow \mathbb{Z}.$$

It is also useful to think of the two-chain $\mathcal{D}(\phi)$, which is gotten as a formal sum of regions in $\Sigma - \alpha_1 - \cdots - \alpha_g - \beta_1 - \cdots - \beta_g$, where a region is counted with multiplicity $n_p(\phi)$, where here $p \in \Sigma$ is any point in this region. Given a Whitney disk, we can consider its space of holomorphic representatives $\mathcal{M}(\phi)$, using the induced complex structure on $\mathrm{Sym}^g(\Sigma)$. If this space is non-empty for all choices of almost-complex structure, then the associated two-chain $\mathcal{D}(\phi)$ has only non-negative local multiplicities. The group $\mathbb{R}$ acts on $\mathcal{M}(\phi)$ by translation. The moduli space $\mathcal{M}(\phi)$ has an expected dimension $\mu(\phi)$, which is obtained as the Fredholm index of the linearized $\bar{\partial}$-operator. This quantity, the *Maslov index*, is denoted $\mu(\phi)$.

It is sometimes necessary to perturb the holomorphic condition to guarantee that moduli spaces are manifolds of the expected dimension. It is useful (though slightly imprecise) to think of a holomorphic disk in $\mathcal{M}(\phi)$ as a pair consisting of a holomorphic surface F with marked boundary, together with a degree g holomorphic projection map π from F to the standard disk, and also a map f from F into Σ.

Here, f maps π^{-1} of the subarc of the boundary of $\mathbb{D}$ with negative resp. positive real part into the subset $\alpha_1 \cup \cdots \cup \alpha_g$ resp $\beta_1 \cup \cdots \cup \beta_g$.

We now consider the complex $CF^-(Y)$ which is the free $\mathbb{Z}[U]$-module generated by $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$, with differential given by

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} \sum_{\{\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y}) \mid \mu(\phi)=1\}} \#\left(\frac{\mathcal{M}(\phi)}{\mathbb{R}}\right) U^{n_w(\phi)} \boldsymbol{y}. \tag{1}$$

In the case where Y is an integral homology sphere, the above sum is readily seen to be finite. (In the case where the first Betti number is positive, some further constraints must be placed on the Heegaard diagram.) With the help of Gromov's compactification of the space pseudo-holomorphic curves [14], one can see that $\partial^2 = 0$.

According to [33], the homology groups $HF^-(Y)$ of $CF^-(Y)$ are a topological invariant of Y. Indeed, the chain homotopy type of $CF^-(Y)$ is a topological invariant, and, since $CF^-(Y)$ is a module over $\mathbb{Z}[U]$, there are a number of other associated constructions. For example, we can form the chain complex $CF^\infty(Y)$ obtained by inverting U, i.e. a chain complex over $\mathbb{Z}[U, U^{-1}]$, with differential as in Equation (1). The quotient of $CF^\infty(Y)$ by $CF^-(Y)$ is a complex $CF^+(Y)$ which is often more convenient to work with. The corresponding homology groups are denoted $HF^\infty(Y)$ and $HF^+(Y)$ respectively. Also, there is a chain complex $\widehat{CF}$ obtained by setting $U = 0$; explicitly, it is generated freely over $\mathbb{Z}$ by $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$, and endowed with the differential

$$\hat{\partial} \boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} \sum_{\{\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y}) \mid \mu(\phi)=1, n_w(\phi)=0\}} \#\left(\frac{\mathcal{M}(\phi)}{\mathbb{R}}\right) \boldsymbol{y},$$

and its homology (also a topological invariant of Y) is denoted $\widehat{HF}(Y)$.

The invariants $HF^-(Y)$, $HF^\infty(Y)$, and $HF^+(Y)$, together with the exact sequence connecting them, are crucial ingredients in the construction of a Heegaard–Floer invariant Φ for closed, smooth four-manifolds. We will say only little more about this invariant here, referring the reader to [27] for its construction.

2. Heegaard–Floer homology of knots

Heegaard–Floer homology for three-manifolds has a refinement to an invariant for null-homologous knots in a three-manifold, as defined in [31], and also independently by Rasmussen in [38].

A knot K in a three-manifold Y is specified by a Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ for Y, together with a pair w and z of basepoints in Σ. The knot K is given as follows. Connect w and z by an arc ξ in $\Sigma - \alpha_1 - \cdots - \alpha_g$ and an arc η in $\Sigma - \beta_1 - \cdots - \beta_g$. The arcs ξ and η are then pushed into U_α and U_β respectively, so that they both meet Σ only at w and z, giving new arcs ξ' and η'. Our knot K, then, is given by

$\xi' - \eta'$. For simplicity, we consider here the case where the ambient manifold Y is the three-sphere S^3.

The new basepoint z gives the Heegaard–Floer complex a filtration. Specifically, we can construct a map

$$F\colon \mathbb{T}_\alpha \cap \mathbb{T}_\beta \longrightarrow \mathbb{Z}$$

by

$$F(\boldsymbol{x}) - F(\boldsymbol{y}) = n_z(\phi) - n_w(\phi), \tag{2}$$

where $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$. It is easy to see that this quantity is independent of the choice of ϕ, depending only on $\boldsymbol{x}$ and $\boldsymbol{y}$. Moreover, if $\boldsymbol{y}$ appears with non-zero multiplicity in $\hat{\partial}(\boldsymbol{x})$, then $F(\boldsymbol{x}) \geq F(\boldsymbol{y})$. This follows from the fact that there is a pseudo-holomorphic disks $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ with $n_w(\phi) = 0$, and also $n_z(\phi) \geq 0$, since a pseudo-holomorphic disks meets the subvariety V_z with non-negative intersection number.

Equation (2) defines F uniquely up to an overall translation. This indeterminacy will be removed presently.

The filtered chain homotopy type of this filtered chain complex is an invariant of the knot K. For example, the homology of the associated graded object, the *knot Floer homology* is an invariant of $K \subset S^3$, defined by

$$\widehat{HFK}(S^3, K) = \bigoplus_{s\in\mathbb{Z}} \widehat{HFK}(S^3, K, s),$$

where $\widehat{HFK}(S^3, K, s)$ is the homology group of the chain complex generated by intersection points $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ with $F(\boldsymbol{x}) = s$, endowed with differential

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y}\in\mathbb{T}_\alpha\cap\mathbb{T}_\beta} \sum_{\left\{\phi\in\pi_2(\boldsymbol{x},\boldsymbol{y}) \;\middle|\; \substack{\mu(\phi)=1,\\ n_w(\phi)=n_z(\phi)=0}\right\}} \#\left(\frac{\mathcal{M}(\phi)}{\mathbb{R}}\right)\boldsymbol{y}.$$

The graded Euler characteristic of this theory is the Alexander polynomial of K, in the sense that

$$\Delta_K(T) = \sum_{s\in\mathbb{Z}} \chi(\widehat{HFK}_*(K, s)) \cdot T^s. \tag{3}$$

This formula can be used to pin down the additive indeterminacy of F: we require that F be chosen so that the graded Euler characteristic is the symmetrized Alexander polynomial. In fact, this symmetry has a stronger formulation, as a relatively graded isomorphism

$$\widehat{HFK}_*(K, s) \cong \widehat{HFK}_*(K, -s). \tag{4}$$

3. Heegaard–Floer homology for links

Heegaard–Floer homology groups of knots can be generalized to the case of links in S^3. For an ℓ-component link, we consider a Heegaard diagram with genus g Heegaard surface, and two $(g + \ell - 1)$-tuples attaching circles $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_{g+\ell-1}\}$

and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_{g+\ell-1}\}$. We require $\{\alpha_1, \ldots, \alpha_{g+\ell-1}\}$ to be disjoint and embedded, and to span a g-dimensional lattice in $H_1(\Sigma; \mathbb{Z})$. The same is required of the $\{\beta_1, \ldots, \beta_{g+\ell-1}\}$. Clearly, $\Sigma - \alpha_1 - \cdots - \alpha_{g+\ell-1}$ consists of ℓ components $A_1, \ldots, A_\ell$. Similarly, $\Sigma - \beta_1 - \cdots - \beta_{g+\ell-1}$ consists of ℓ components $B_1, \ldots, B_\ell$. We assume that this Heegaard diagram has the special property that $A_i \cap B_i$ is non-empty. Indeed, for each $i = 1, \ldots, \ell$, we choose basepoints w_i and z_i to lie inside $A_i \cap B_i$. We call the collection of data $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{w_1, \ldots, w_\ell\}, \{z_1, \ldots, z_\ell\})$ a *2ℓ-pointed Heegaard diagram*.

A link can now be constructed in the following manner. Connect w_i and z_i by an arc ξ_i in A_i and an arc η_i in B_i. Again, the arc ξ_i resp. η_i is pushed into U_α resp. U_β to give rise to a pair of arcs ξ_i' and η_i'. The link L is given by $\cup_{i=1}^{\ell} \xi_i' - \eta_i'$. For a 2ℓ-pointed Heegaard diagram for S^3 $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{w_1, \ldots, w_\ell\}, \{z_1, \ldots z_\ell\})$, if L is the link obtained in this manner, we say that the Heegaard diagram is *compatible* with the link L.

We will need to make an additional assumption on the Heegaard diagram. A *periodic domain* is a two-chain in Σ of the form

$$\sum c_i(A_i - B_i),$$

where $c_i \in \mathbb{Z}$. Our assumption is that all non-zero periodic domains have some positive and some negative local multiplicities c_i. This assumption on the pointed Heegaard diagram is called *admissibility*.

Let $L \subset S^3$ be an ℓ-component link, suppose that L is embedded so that the restriction of the height function to L has b local maxima, then we can construct a compatible 2ℓ-pointed Heegaard diagram with Heegaard genus $g = b - \ell$.

For example, consider the two-component "Conway link" pictured in Figure 1. This is the $(2, -3, -2, 3)$ pretzel link, also known as L10n59 in Thistlethwaite's link table [1]. For this link, $b = 4$, and hence we can draw it on a surface of genus

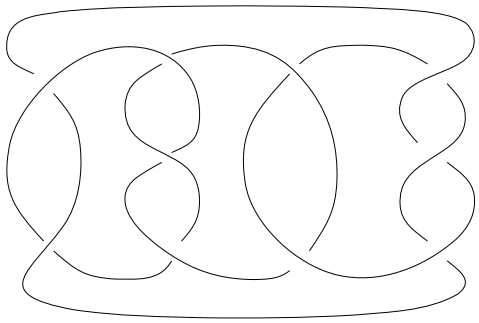


Figure 1. The Conway link.

$g = 2$, as illustrated in Figure 2. It is straightforward to verify that the space of periodic domains is one-dimensional; drawing a picture of this generator, it is also straightforward to see that the diagram is admissible.

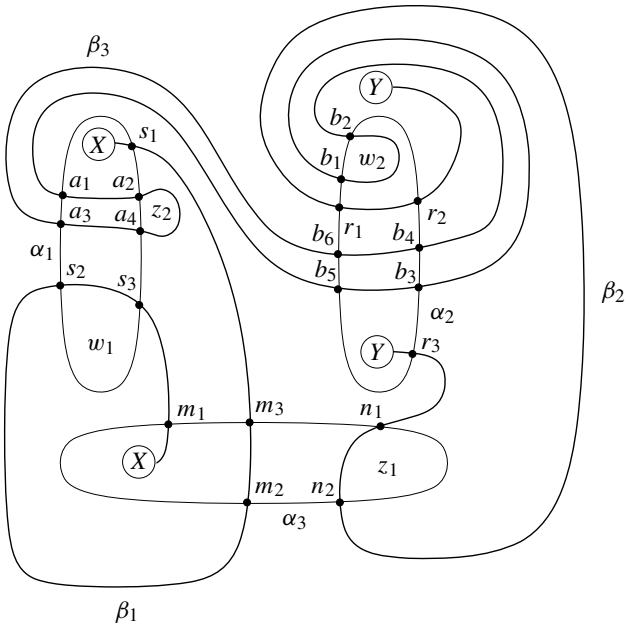


Figure 2. Pointed Heegaard diagram for the Conway link. This picture takes place on the genus two surface obtained by identifying the two disks labeled by X and the two disks labeled by Y.

Now, we work inside $\mathrm{Sym}^{g+\ell-1}(\Sigma)$, relative to the tori $\mathbb{T}_\alpha = \alpha_1 \times \cdots \times \alpha_{g+\ell-1}$ and $\mathbb{T}_\beta = \beta_1 \times \cdots \times \beta_{g+\ell-1}$, and consider intersection points of $\mathbb{T}_\alpha$ and $\mathbb{T}_\beta$; i.e. $g+\ell-1$-tuples of points $(x_1, \ldots, x_{g+\ell-1})$ with $x_i \in \alpha_i \cap \beta_{\sigma(i)}$ for some permutation σ in the symmetric group on $g+\ell-1$ letters. We then form the chain complex $\widehat{CFL}(S^3, L)$ generated freely by these intersection points.

For example, for the figure illustrated in Figure 2, the curves α_i and β_j intersect according to the pattern

$\cap$	α_1	α_2	α_3
β_1	$\{s_1, s_2, s_3\}$	$\emptyset$	$\{m_1, m_2, m_3\}$
β_2	$\emptyset$	$\{r_1, r_2, r_3\}$	$\{n_1, n_2\}$
β_3	$\{a_1, \ldots, a_4\}$	$\{b_1, \ldots, b_6\}$	$\emptyset$

Now, there are exactly two permutations of $\{1, 2, 3\}$ for which $\alpha_i \cap \beta_{\sigma(i)}$ is non-trivial for all i. This gives two types of intersection points of $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$, namely, $a_i \times m_j \times r_k$ (with $i = 1, \ldots, 4$, $j = 1, \ldots, 3$, $k = 1, \ldots, 3$) and also $b_i \times n_j \times s_k$ (with $i = 1, \ldots, 6$, $j = 1, 2$, $k = 1, 2, 3$). This gives a chain complex with a total of 72 generators.

The complex $\widehat{CFL}$ has a grading, the *Maslov grading*, which is specified up to overall translation by the convention

$$\mathrm{gr}(\boldsymbol{x}) - \mathrm{gr}(\boldsymbol{y}) = \mu(\phi) - 2\sum_{i=1}^{\ell} n_{w_i}(\phi),$$

where $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ is any Whitney disk connecting $\boldsymbol{x}$ and $\boldsymbol{y}$. The parity of the Maslov grading depends on the local sign of the intersection number of $\mathbb{T}_\alpha$ and $\mathbb{T}_\beta$ at $\boldsymbol{x}$.

But $\widehat{CFL}$ has an additional grading, the $\mathbb{H}$-grading. To define this, we associate to each $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ the pair of vectors

$$n_{\boldsymbol{w}}(\phi) = (n_{w_1}(\phi), \ldots, n_{w_\ell}(\phi)) \quad \text{and} \quad n_{\boldsymbol{z}}(\phi) = (n_{z_1}(\phi), \ldots, n_{z_\ell}(\phi)).$$

We have a function $F\colon \mathbb{T}_\alpha \cap \mathbb{T}_\beta \longrightarrow \mathbb{Z}^\ell \cong H_1(S^3 - L; \mathbb{Z})$ (where the latter identification is given by the meridians of the link L) specified uniquely up to translation by the formula

$$F(\boldsymbol{x}) - F(\boldsymbol{y}) = n_{\boldsymbol{z}}(\phi) - n_{\boldsymbol{w}}(\phi),$$

where ϕ is any choice of homotopy class in $\pi_2(\boldsymbol{x}, \boldsymbol{y})$.

Endow $\widehat{CFL}(S^3, L)$ with the differential

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} \sum_{\left\{\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y}) \;\middle|\; \substack{\mu(\phi) = 1, \\ n_{\boldsymbol{w}}(\phi) = n_{\boldsymbol{z}}(\phi) = 0}\right\}} \#\left(\frac{\mathcal{M}(\phi)}{\mathbb{R}}\right) \boldsymbol{y}.$$

It is easy to see that this differential drops Maslov grading by one. Moreover, the complex naturally splits into summands indexed by elements of $\mathbb{Z}^\ell \cong H_1(S^3 - L; \mathbb{Z})$ specified by the function F. We find it natural to think of these summands, in fact, as indexed by the affine lattice $\mathbb{H} \subset H_1(S^3 - L; \mathbb{R})$ over $H_1(S^3 - L; \mathbb{Z})$, given by elements

$$\sum_{i=1}^{\ell} a_i \cdot [\mu_i],$$

where $a_i \in \mathbb{Q}$ satisfies the property that

$$2a_i + \mathrm{lk}(L_i, L - L_i)$$

is an even integer. The translational ambiguity of the map is then pinned down by the following generalization of Equation (4):

$$\widehat{HFL}_*(\vec{L}, h) \cong \widehat{HFL}_*(\vec{L}, -h). \tag{5}$$

In practice, it is easy to calculate the difference in F for any two intersection of $\mathbb{T}_\alpha$ and $\mathbb{T}_\beta$ which have the same type (i.e. same pattern of intersection $\alpha_i \cap \beta_{\sigma(i)}$). To this end, it suffices to find for each pair of intersection points $x, x' \in \alpha_i \cap \beta_j$, a disk (or more generally a compact surface with a single boundary component) in Σ which

meets α_i along one arc in its boundary and β_j along the complementary arc, carrying the intersection points of the closures of the arcs to x and x'. We then define the "relative difference" of x and x', $F^{i,j}(x) - F^{i,j}(x')$, to be $n_{\boldsymbol{z}} - n_{\boldsymbol{w}}$ for this disk (or surface). It is easy to see then that if $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ are two intersection points with the same type (as specified by σ), then

$$F(\boldsymbol{x}) - F(\boldsymbol{y}) = \sum_{i=1}^{g+\ell-1} F^{i,\sigma(i)}(x_i) - F^{i,\sigma(i)}(y_i),$$

where $\boldsymbol{x} = (x_1, \dots, x_{g+\ell-1})$ and $y = (y_1, \dots, y_{g+\ell-1})$. This determines $F(\boldsymbol{x}) - F(\boldsymbol{y})$ for $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same type. Different types can then be compared by choosing homotopy classes connecting them (and in suitable circumstances, the translational ambiguity can be removed using Equation (5)).

We display the relative differences for the various intersection points for the diagram from Figure 2.

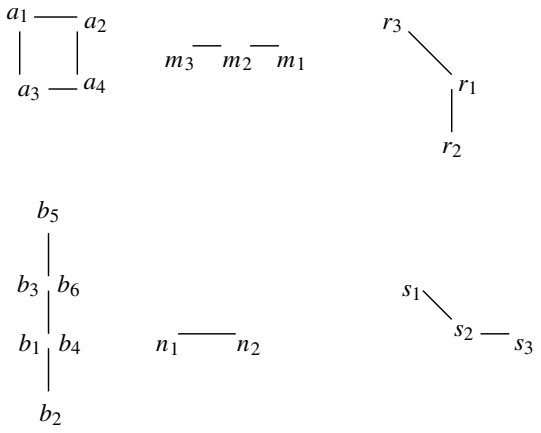


Figure 3. Generators for $\widehat{HFL}$ of the Conway link. We illustrate the relative differences of the various intersection points of α_i and β_j. A horizontal resp. vertical segment denotes two intersection points whose relative difference is one in the first resp. second component; e.g. there is a disk in Figure 2 ϕ from a_2 to a_1 with $n_z - n_w$ given by $(1, 0)$, while there is one from b_5 to b_3 with relative difference given by $(0, 1)$. Finally, for the diagonal edges, we have a disk from r_3 to r_1 with relative difference $(-1, 1)$.

It is now straightforward to verify that the ranks of the chain groups in each value of F is given as in Figure 4. It is more challenging to calculate the homology groups of $\widehat{CFL}$.

Some aspects are immediate. For example, it follows by glancing at Figure 4, and comparing Equation (5) that the homology in $\mathbb{H}$-grading $(-2, 2)$ is trivial (as there are no generators in the $\mathbb{H}$-grading $(2, -2)$), and that in $\mathbb{H}$-gradings $(-1, 2)$ and $(-2, 1)$ the groups $\widehat{HFL}$ have rank one. This already suffices to determine the convex hull

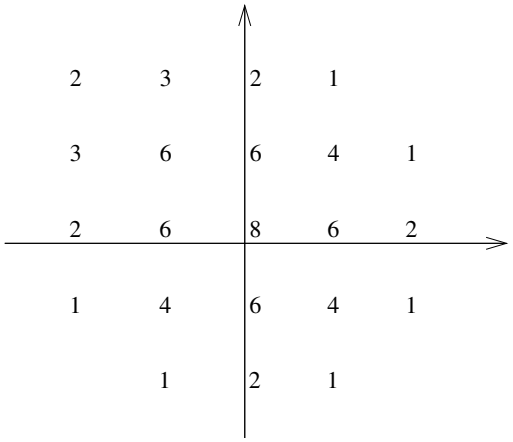


Figure 4. Generators for $\widehat{CFL}$ of the Conway link. The 72 generators of $\widehat{CFL}$ coming from the diagram in Figure 2 are partitioned into various filtration levels. In this figure, each integer represents the number of generators in the filtration level specified by its coordinates in the plane.

of $h \in \mathbb{H}$ for which $\widehat{HFL}(L, h)$ is non-trivial, as required for the application to the Thurston norm below (see esp. Equation (7)).

Also, the calculation of $\widehat{HFL}(L, (x, y))$ with $(x, y) \in \{(0, \pm 2), (\pm 2, 0)\}$ follows from the fact that for each of these $\mathbb{H}$-gradings, every generator has the same Maslov grading.

Next, consider the part in $\mathbb{H}$-grading $(1, 1)$. There are four generators

$$a_1 \times m_1 \times r_1, \quad a_2 \times m_2 \times r_1, \quad a_4 \times r_3 \times m_1, \quad b_5 \times n_1 \times s_3.$$

For the case where $\boldsymbol{x} = b_5 \times n_1 \times s_3$ and $\boldsymbol{y} \in \{a_1 \times m_1 \times r_1, a_2 \times m_2 \times r_1\}$, there is a homotopy class $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ whose associated two-chain $\mathcal{D}(\phi)$ is a hexagon. For the case where $\boldsymbol{y} = a_2 \times m_2 \times r_1$, we illustrate this in Figure 5.

A hexagon gives rise to a flow-line connecting $\boldsymbol{x}$ to $\boldsymbol{y}$. To this end, we think of a holomorphic disk in $\mathrm{Sym}^3(\Sigma)$ as a branched triple-cover F of the standard disk, together with a map of F into Σ. The given hexagonal domain in Σ can be realized as a branched triple-cover of the disk $\mathbb{D}$. Moreover, any other domain connecting $\boldsymbol{x}$ to $\boldsymbol{y}$ has negative local multiplicity somewhere. Hence, we have that

$$\partial b_5 \times n_1 \times s_3 = a_1 \times m_1 \times r_1 + a_2 \times m_2 \times r_1.$$

It can also be seen that

$$\mathrm{gr}(b_5 \times n_1 \times s_3) = \mathrm{gr}(a_4 \times m_1 \times r_3),$$

but there are no non-negative domains from $a_4 \times m_1 \times r_3$ to either of $\{a_1 \times m_1 \times r_1, a_2 \times m_2 \times r_1\}$. It follows at once that $\widehat{HFL}(L, (1, 1))$ has rank two.

With some additional work, one can verify that all the link Floer homology groups of the Conway link are as displayed below in Figure 6.

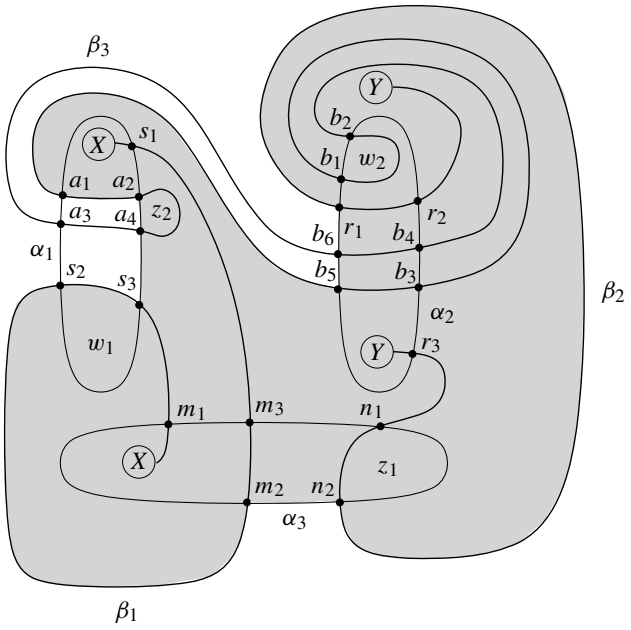


Figure 5. A flowline. The complement of the shaded region gives a hexagon connecting $b_5 \times n_1 \times s_3$ to $a_2 \times m_2 \times r_1$.

4. Basic properties

Perhaps the single most fundamental property of Heegaard–Floer homology is that it satisfies an exact triangle for surgeries. More precisely, a *triad* of three-manifolds Y_1, Y_2, Y_3 is a cyclically ordered triple of three-manifolds obtained as follows. Let M be a three-manifold with torus boundary, and fix three simple, closed curves in its boundary γ_1, γ_2, γ_3 any two of which intersect transversally in one point, and ordered so that there are orientations on the three curves so that

$$\#(\gamma_1 \cap \gamma_2) = \#(\gamma_2 \cap \gamma_3) = \#(\gamma_3 \cap \gamma_1) = -1.$$

We let Y_i be the three-manifold obtained by Dehn filling M along the curve γ_i.

Theorem 4.1. *Let Y_1, Y_2, and Y_3 be a triad of three-manifolds. Then, there is an long exact sequence of the form*

$$\cdots \longrightarrow \widehat{HF}(Y_1) \longrightarrow \widehat{HF}(Y_2) \longrightarrow \widehat{HF}(Y_3) \longrightarrow \cdots.$$

The maps in the exact triangle are induced by the three natural two-handle cobordisms connecting Y_i and Y_{i+1} (where we i as an integer modulo 3), in a manner made precise in [27].

The above surgery exact sequence is similar to an exact sequence established by Floer for instanton Floer homology (only using a restricted class of triads) [9], [2]. An

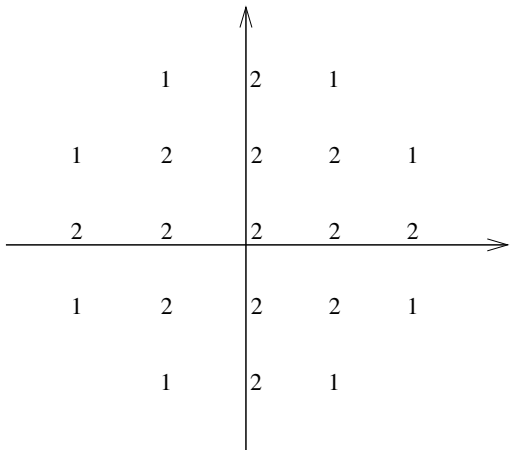


Figure 6. Ranks of $\widehat{HFL}$ for the Conway link. The ranks are displayed, along with their $\mathbb{H}$ grading, thought of as coordinates in the plane.

analogous exact sequence has been established for Floer homology of Seiberg–Witten monopoles, see [20]. There are also related exact sequences in symplectic geometry, cf. [40].

Note that there are other variants of the exact triangle, and indeed, there are certain other related calculational techniques for Heegaard–Floer homology. For example, for a knot K in an integral homology sphere Y, the filtered chain homotopy type of the induced knot invariant can be used to calculate the Floer homology groups of arbitrary surgeries on K, [28].

5. Three applications

Heegaard–Floer homology is particularly well suited to problems in knot-theory and three-manifold topology which can be formulated in terms of the existence of four-dimensional cobordisms. We focus here on a few concrete problems which can be formulated for knots and links in the three-sphere. For some other applications, see [36], [34], [22], [37].

5.1. Thurston norm. Let $K \subset S^3$ be a knot. The *Seifert genus* of K, denoted $g(K)$, is the minimal genus of any embedded surface $F \subset S^3$ with boundary K. Clearly, if $g(K) = 0$, then K is the unknot.

According to [30], knot Floer homology detects the Seifert genus of a knot, by the property that

$$g(K) = \max\{s \mid \widehat{HFK}(K, s) \neq 0\}. \tag{6}$$

There is a natural generalization of the knot genus and indeed of Equation (6). This is best formulated in terms of Thurston's norm on second homology.

Recall that if F is a compact, oriented, but possibly disconnected surface-with-boundary $F = \bigcup_{i=1}^{n} F_i$, its *complexity* is given by

$$\chi_-(F) = \sum_{\{F_i \mid \chi(F_i)\leq 0\}} -\chi(F_i).$$

Given any homology class $h \in H_2(S^3, L)$, it is easy to see that there is a compact, oriented surface-with-boundary embedded in $S^3 - \mathrm{nd}(K)$ representing h. Consider the function from $H^1(S^3 - L; \mathbb{Z})$ to the integers defined by

$$x(h) = \min_{\{F\hookrightarrow S^3-\mathrm{nd}(K) \mid [F]=\mathrm{PD}[h]\}} \chi_-(F).$$

Indeed, according to Thurston [41], this function x satisfies an inequality $x(h_1+h_2) \leq x(h_1) + x(h_2)$, and it is linear on rays, i.e. given $h \in H_2(S^3, L)$ and a non-negative integer n, we have that $x(n \cdot h) = nx(h)$. Thus, x can be naturally extended to a semi-norm on $H^1(S^3 - L; \mathbb{R})$, the *Thurston semi-norm*. In fact, this semi-norm is uniquely specified by its unit ball

$$B_x = \{h \in H^1(S^3 - L; \mathbb{R}) \mid x(h) \leq 1\},$$

which is a polytope $H^1(S^3-L; \mathbb{R})$ whose vertices lie at lattice points in $H^1(S^3-L; \mathbb{Z})$.

Equation (6) can now be generalized as follows. A *trivial component* of a link L is a component $K \subset L$ which is unknotted an geometrically unlinked from the complement $L - K$. Suppose that L is a link with no trivial components. Then, given $s \in H^1(S^3 - L)$,

$$x(h) + \sum_{i=1}^{\ell} |\langle \mu_i, h\rangle| = 2 \cdot \max_{\{s\in\mathbb{H} \mid \widehat{HFL}(L,s)\neq 0\}} \langle h, s\rangle, \tag{7}$$

where here $\langle , \rangle$ denotes the Kronecker pairing of $H_1(S^3 - L; \mathbb{R})$ with $H^1(S^3 - L; \mathbb{R})$.

This formula can be thought of more geometrically from the following point of view. Consider the dual norm $x^*\colon H_1(S^3 - L; \mathbb{R}) \longrightarrow \mathbb{R}$ given by

$$x^*(s) = \max_{\{h\in B_x\}} \langle s, h\rangle.$$

The unit ball B_{x*} is a (possibly degenerate) polytope in $H_1(S^3 - L; \mathbb{R})$ called the *dual Thurston polytope*. Equation (7) states that for a link L with no trivial components, if we take the convex hull of the set of $s \in \mathbb{H}$ with $\widehat{HFL}(L, s)$, and rescale that polytope by a factor of two, then we obtain the sum of the dual Thurston polytope with the symmetric hypercube in $H_1(S^3 - L; \mathbb{R})$ with edge-length two.

Of course, the Thurston norm can be defined for closed three-manifolds, as well. In fact, a result analogous to Equation (7) can be proved for closed three-manifolds Y,

instead of link complements. An analogous result has been shown to hold for Seiberg–Witten monopole Floer homology [19] (but at present there is no analogue of knot and link Floer homology in gauge-theoretic terms).

Although the statement of Equation (6) does not explicitly involve any four-dimensional theory, the proof of this result does use the full force of Heegaard–Floer homology, combined with Gabai's theory of sutured manifolds, and recent results in symplectic geometry. Specifically, according to a combination of theorems of Gabai [11], [12], Eliashberg–Thurston [6], and a result of Eliashberg [5] and independently Etnyre [7], if $K \subset S^3$ is a knot of genus g, then its zero-surgery $S^3_0(K)$ separates a symplectic manifold. Non-vanishing theorems for the Heegaard–Floer invariant Φ for symplectic four-manifolds, which in turn are built on the symplectic Lefshetz pencils of Donaldson [3], then give a non-vanishing result for the Heegaard–Floer $S^3_0(K)$ from which Equation (6) follows.

These results can be further generalized to give Equation (7). Specifically, an n-component link in S^3 naturally gives rise to a connected knot in the $(n-1)$-fold connected sum of $S^2 \times S^1$. A genus bound analogous to Equation (6) has been shown by Ni in [25], which in effect establishes Equation (7), in the case where h is one of the 2^ℓ cohomology classes whose evaluation on each meridian for L has absolute value equal to one. Equation (7) then follows from the manner in which the Thurston norm and link Floer homology transform under cabling, see also [16].

As an illustration of Equation (6), consider the Conway link from Figure 1. According to the calculations displayed in Figure 6, together with this equation, we conclude that the dual Thurston polytope of the Conway link is as illustrated in Figure 7.

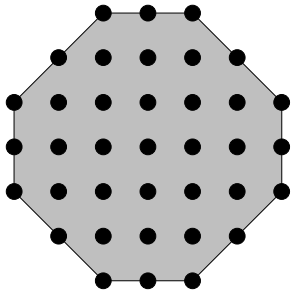


Figure 7. Dual Thurston polytope for the Conway link. Lattice points in $H_1(S^3 - L; \mathbb{Z})$ are indicated by solid circles.

Figure 7 suggests that there are surfaces F_1 the complement of the Conway link L, with ∂F_1 consisting of a longitude belonging to a first component of L, and some number of copies of the meridian of the second component, and also with $\chi_-(F_1) = 5$. In fact, such a surface can be easily obtained by puncturing a genus one Seifert surface for one of the trefoil components in two additional points. A similar surface can be found for the other component of L.

Note that a verification of the dual Thurston polytope for the Conway link can be easily obtained by more classical methods (cf. [23]); however, the computation given here is fairly easy (and hopefully illustrates the theory).

5.2. Slice genus. A *slice surface* for a knot K is a smoothly embedded surface-with-boundary $F \subset B^4$ which meets S^3 along its boundary, which is K. The *slice genus* of a knot $g_*(K)$ is the minimal genus of any slice surface for K. Heegaard–Floer homology can be used to give information about this quantity, as follows.

Recall that knot Floer homology is the homology of an associated graded object which is induced by the filtration of a chain complex which calculates $\widehat{HF}(S^3) \cong \mathbb{Z}$. But the entire filtered chain homotopy type of the complex is a knot invariant. Denote the sequence of subcomplexes $F_i \subset F_{i+1}$, so that for all sufficiently small integers i, $F_i = 0$, while for all sufficiently large integers, $F_i = \widehat{CF}(S^3)$. There is another natural invariant which can be associated to a knot, which is the minimal i for which the map $H_*(F_i) \longrightarrow \widehat{HF}(S^3)$ is non-trivial. According to [29] and independently [38],

$$|\tau(K)| \leq g^*(K).$$

Intriguingly, Rasmussen [39] has shown that a very similar algebraic construction on Khovanov's homology [17], [21], can be used to define a similar (but entirely combinatorial) numerical invariant $s(K)$. Although both $\tau(K)$ and $s(K)$ share many formal properties (and hence agree on many knots), Hedden and Ording have recently shown [15] that these two invariants are in fact distinct. Their examples are certain twisted Whitehead doubles of the trefoil.

5.3. Unknotting numbers. The *unknotting number* $u(K)$ is the minimal number of crossing changes required to transform K into an unknot. An n-step unknotting of a knot K in effect gives an immersed disk in B^4 with n double-points. Resolving these double-points, we obtain a slice surface for K with genus n. This observation immediately verifies the inequality

$$g^*(K) \leq u(K).$$

For some classes of knots, these two quantities are equal. For example, for the (p, q) torus knot, $g^*(K) = u(K) = (p-1)(q-1)/2$. This was first shown by Kronheimer and Mrowka in [18] (though it has alternative proofs now using either τ [29] or the Khovanov–Rasmussen invariant s [39]).

But there are Floer-theoretic bounds on $u(K)$ which go beyond the slice genus, cf. [35], [26].

Suppose that K has $u(K) = 1$. Then, Montesinos observed [24] that the branched double-cover of S^3 with branching locus K, denoted $\Sigma(K)$, can be realized as $\pm d/2$-surgery on a different knot $C \subset S^3$, where here $d = |\Delta_K(-1)|$. Obstructions to this can sometimes be given using Heegaard–Floer homology.

To do this in a useful manner, we must understand first $\widehat{HF}(\Sigma(K))$. For some knots, this is a straightforward matter. For example, when K is a knot which admits an alternating projection, an easy induction using Theorem 4.1 shows that $\widehat{HF}(\Sigma(K))$ is a free $\mathbb{Z}$-module of rank $|\Delta_K(-1)|$. This means that the Heegaard–Floer homology groups of these three-manifolds is as simple as possible. For any rational homology three-sphere Y (i.e. closed three-manifold with $H_1(Y;\mathbb{Q}) = 0$), the Euler characteristic of $\widehat{HF}(Y)$ is $|H_1(Y;\mathbb{Z})|$, the number of elements in $H_1(Y;\mathbb{Z})$. A rational homology three-sphere whose homology group $\widehat{HF}(Y)$ is a free module of rank $|H_1(Y;\mathbb{Z})|$ is called an *L-space*. Thus, if K is a knot with alternating projection, then $\Sigma(K)$ is an L-space.

There are obstructions to realizing an L-space as surgery on a knot in S^3, These obstructions are phrased in terms of an additional $\mathbb{Q}$-grading on the Heegaard–Floer homology [35], analogous to an invariant defined by Frøyshov [10] in the context of Seiberg–Witten theory. Moreover, this $\mathbb{Q}$-grading can be explicitly calculated for $\Sigma(K)$ for an alternating knot K from its Goeritz matrix. Rather than stating these results precisely, we content ourself here with including a picture of an eight-crossing alternating knot (8_{10}, see Figure 8) whose unknotting number can be shown to equal two via these (and presently, no other known) techniques.

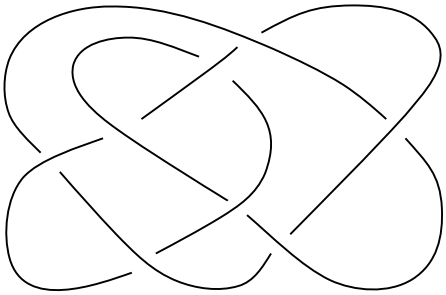


Figure 8. A knot with $u = 2$.

Combining these obstructions with recent work of Gordon and Luecke [13], one can classify all knots with 10 and fewer crossings which have unknotting number equal to one. Indeed, a different application of Heegaard–Floer homology along similar lines discovered by Owens [26] can be used to complete the unknotting number table for prime knots with nine or fewer crossings.

References

[1] Bar-Natan, D., Thistlethwaite's link table. http://katlas.math.toronto.edu/wiki/.

[2] Braam, P., and Donaldson, S. K., Floer's work on instanton homology, knots, and surgery. In *The Floer Memorial Volume* (ed. by H. Hofer, C. H. Taubes, A. Weinstein, and E. Zehnder), Progr. Math. 133, Birkhäuser, Basel 1995, 195–256.

[3] Donaldson, S. K., An application of gauge theory to four-dimensional topology. *J. Differential Geom.* **18** (2) (1983), 279–3153.

[4] Donaldson, S. K., *Floer homology groups in Yang-Mills theory*. With the assistance of M. Furuta and D. Kotschick, Cambridge Tracts in Math. 147, Cambridge University Press, Cambridge 2002.

[5] Eliashberg, Y. M., Few remarks about symplectic filling. *Geom. Topol.* **8** (2004), 277–293.

[6] Eliashberg, Y. M., and Thurston, W. P., *Confoliations*. Univ. Lecture Ser. 13, Amer. Math. Soc., Providence, RI, 1998.

[7] Etnyre, J. B., On symplectic fillings. *Algebr. Geom. Topol.* **4** (2004), 73–80.

[8] Floer, A., Morse theory for Lagrangian intersections. *J. Differential Geom.* **28** (1988), 513–547.

[9] Floer, A., Instanton homology and Dehn surgery. In *The Floer Memorial Volume* (ed. by H. Hofer, C. H. Taubes, A. Weinstein, and E. Zehnder), Progr. Math. 133, Birkhäuser, Basel 1995, 77–97.

[10] Frøyshov, K. A., The Seiberg-Witten equations and four-manifolds with boundary. *Math. Res. Lett* **3** (1996), 373–390.

[11] Gabai, D., Foliations and the topology of 3-manifolds. *J. Differential Geom.* **18** (3) (1983), 445–503.

[12] Gabai, D., Foliations and the topology of 3-manifolds III. *J. Differential Geom.* **26** (3) (1987), 479–536.

[13] Gordon, McA. C., and Luecke, J., Knots with unknotting number 1 and essential Conway spheres. math.GT/0601265.

[14] Gromov, M., Pseudo holomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), 307–347.

[15] Hedden, M., The Ozsváth-Szabó and Rasmussen concordance invariants are not equal. math.GT/0512348.

[16] Hedden, M., On knot Floer homology and cabling. *Algebr. Geom. Topol.* **5** (2005), 1197–1222.

[17] Khovanov, M., A categorification of the Jones polynomial. *Duke Math. J.* **101** (3) (2000), 359–426.

[18] Kronheimer, P. B., and Mrowka, T. S., Gauge theory for embedded surfaces. I. *Topology* **32** (4) (1993), 773–826.

[19] Kronheimer, P. B., and Mrowka, T. S., Scalar curvature and the Thurston norm. *Math. Res. Lett.* **4** (6) (1997), 931–937.

[20] Kronheimer, P. B., Mrowka, T. S., Ozsváth, P. S., and Szabó, Z., Monopoles and lens space surgeries. math.GT/0310164.

[21] Lee, E.-S., An endomorphism of the Khovanov invariant. *Adv. Math.* **197** (2) (2005), 554–586.

[22] Lisca, P., and Stipsicz, A. I., Ozsváth-Szabó invariants and tight contact three-manifolds. I. *Geom. Topol.* **8** (2004), 925–945.

[23] McMullen, C. T., The Alexander polynomial of a 3-manifold and the Thurston norm on cohomology. *Ann. Sci. École Norm. Sup.* **35** (2) (2002), 153–171.

[24] Montesinos, J. M., Surgery on links and double branched covers of S^3. In *Knots, groups, and 3-manifolds* (Papers dedicated to the memory of R. H. Fox), Ann. of Math. Stud. 84, Princeton University Press, Princeton, N.J., 1975, 227–259.

[25] Ni, Y., A note on knot Floer homology of links. math.GT/0506208.

[26] Owens, B., Unknotting information from Heegaard Floer homology. math.GT/0506485.

[27] Ozsváth, P. S., and Szabó, Z., Holomorphic triangles and invariants for smooth four-manifolds. math.SG/0110169.

[28] Ozsváth, P. S., and Szabó, Z., Knot Floer homology and rational surgeries. math.GT/0504404.

[29] Ozsváth, P. S., and Szabó, Z., Knot Floer homology and the four-ball genus. *Geom. Topol.* **7** (2003), 615–643.

[30] Ozsváth, P. S., and Szabó, Z., Holomorphic disks and genus bounds. *Geom. Topol.* **8** (2004), 311–334.

[31] Ozsváth, P. S., and Szabó, Z., Holomorphic disks and knot invariants. *Adv. Math.* **186** (1) (2004), 58–116.

[32] Ozsváth, P. S., and Szabó, Z., Holomorphic disks and three-manifold invariants: properties and applications. *Ann. of Math.* (2) **159** (3) (2004), 1159–1245.

[33] Ozsváth, P. S., and Szabó, Z., Holomorphic disks and topological invariants for closed three-manifolds. *Ann. of Math.* (2) **159** (3) (2004), 1027–1158.

[34] Ozsváth, P. S., and Szabó, Z., Heegaard Floer homology and contact structures. *Duke Math. J.* **129** (1) (2005), 39–61.

[35] Ozsváth, P. S., and Szabó, Z., Knots with unknotting number one and Heegaard Floer homology. *Topology* **44** (4) (2005), 705–745.

[36] Ozsváth, P. S., and Szabó, Z., On knot Floer homology and lens space surgeries. *Topology* **44** (6) (2005), 1281–1300.

[37] Rasmussen, J., Lens space surgeries and a conjecture of Goda and Teragaito. *Geom. Topol.* **8** (2004), 1013–1031.

[38] Rasmussen, J. A., Floer homology and knot complements. PhD thesis, Harvard University, 2003; math.GT/0306378.

[39] Rasmussen, J. A., Khovanov homology and the slice genus. math.GT/0402131, 2004.

[40] Seidel, P., A long exact sequence for symplectic Floer cohomology. *Topology* **42** (5) (2003), 1003–1063.

[41] Thurston, W. P., A norm for the homology of 3-manifolds. *Mem. Amer. Math. Soc.* **59** (1986), 99–130.

Department of Mathematics, Columbia University, New York, NY 10027, U.S.A.
E-mail: petero@math.columbia.edu

Department of Mathematics, Princeton University, New Jersey 08544, U.S.A.
E-mail: szabo@math.princeton.edu

The cohomology of automorphism groups of free groups

Karen Vogtmann*

Abstract. There are intriguing analogies between automorphism groups of finitely generated free groups and mapping class groups of surfaces on the one hand, and arithmetic groups such as $\mathrm{GL}(n, \mathbb{Z})$ on the other. We explore aspects of these analogies, focusing on cohomological properties. Each cohomological feature is studied with the aid of topological and geometric constructions closely related to the groups. These constructions often reveal unexpected connections with other areas of mathematics.

Mathematics Subject Classification (2000). Primary 20F65; Secondary, 20F28.

Keywords. Automorphism groups of free groups, Outer space, group cohomology.

1. Introduction

In the 1920s and 30s Jakob Nielsen, J. H. C. Whitehead and Wilhelm Magnus invented many beautiful combinatorial and topological techniques in their efforts to understand groups of automorphisms of finitely-generated free groups, a tradition which was supplemented by new ideas of J. Stallings in the 1970s and early 1980s. Over the last 20 years mathematicians have been combining these ideas with others motivated by both the theory of arithmetic groups and that of surface mapping class groups. The result has been a surge of activity which has greatly expanded our understanding of these groups and of their relation to many areas of mathematics, from number theory to homotopy theory, Lie algebras to bio-mathematics, mathematical physics to low-dimensional topology and geometric group theory.

In this article I will focus on progress which has been made in determining cohomological properties of automorphism groups of free groups, and try to indicate how this work is connected to some of the areas mentioned above. The concept of assigning cohomology groups to an abstract group was originally motivated by work of Hurewicz in topology. Hurewicz proved that the homotopy type of a space with no higher homotopy groups (an *aspherical space*) is determined by the fundamental group of the space, so the homology groups of the space are in fact invariants of the group. Low-dimensional homology groups were then found to have interpretations in terms of algebraic invariants such as group extensions and derivations which had long been studied by algebraists, and a purely algebraic definition of group cohomology

*The author was partially supported by grants from the National Science Foundation

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

was also introduced. These were the beginning steps of a rich and fruitful interaction between topology and algebra via cohomological methods.

Borel and Serre studied the cohomology of arithmetic and S-arithmetic groups by considering their actions on homogeneous spaces and buildings. Thurston studied surface mapping class groups by considering their action on the Teichmüller space of a surface, and this same action was used later by Harer to determine cohomological properties of mapping class groups. Outer automorphism groups of free groups are neither arithmetic groups nor surface mapping class groups, but they have proved to share many algebraic features with both classes of groups, including many cohomological properties. The analogy is continually strengthened as more and more techniques from the arithmetic groups and mapping class groups settings are adapted to the study of automorphism groups of free groups. The adaptation is rarely straightforward, and often serves more as a philosophy than a blueprint. The connection is more than strictly empirical and philosophical, however, due to the natural maps

$$\mathrm{Out}(F_n) \to \mathrm{GL}(n, \mathbb{Z})$$

and

$$\Gamma_{g,s} \to \mathrm{Out}(F_{2g+s-1}).$$

The first map is induced by the abelianization map $F_n \to \mathbb{Z}^n$; it is always surjective and is an isomorphism for $n = 2$. In the second map, the group $\Gamma_{g,s}$ is the mapping class group of a surface of genus g with $s > 0$ punctures, which may be permuted. The map is defined using the observation that a homeomorphism of a surface induces a map on the (free) fundamental group of the surface; it is always injective and is an isomorphism for $g = s = 1$.

2. Outer space and homological finiteness results

In order to adapt techniques Borel and Serre used for arithmetic groups, and those Thurston and Harer used for mapping class groups to the context of automorphism groups of free groups, the first thing one needs is a replacement for the homogeneous or Teichmüller space. A suitable space $\mathcal{O}_n$, now called *Outer space*, was introduced by Culler and Vogtmann in 1986 [12]. The most succinct definition of Outer space is that it is the space of homothety classes of minimal free simplicial actions of $\mathrm{Out}(F_n)$ on $\mathbb{R}$-trees. (Here an $\mathbb{R}$-*tree* is a metric space with a unique arc, isometric to an interval of $\mathbb{R}$, joining any two points and actions are by isometries; an action is *simplicial* if every orbit is discreet and *minimal* if there is no proper invariant subtree.) The topology on the space can be taken to be the equivariant Gromov–Hausdorff topology, or it can be topologized as a space of projective length functions on F_n. Pre-composing an action with an element of $\mathrm{Out}(F_n)$ gives a new action, and this defines the action of $\mathrm{Out}(F_n)$ on the space. This description is efficient, but it is often easier to visualize and to work with Outer space when it is presented instead in terms of *marked graphs*.

2.1. Marked graphs. The quotient of a free action of F_n on a simplicial $\mathbb{R}$-tree is a finite graph with fundamental group F_n and a metric determined by the lengths of the edges. The action is minimal if and only if the quotient graph has no univalent or bivalent vertices. If we have a specific identification of F_n with the fundamental group of the graph then the tree, with its F_n-action, can be recovered as the universal cover of the graph. Thus another way to describe a point in Outer space is to fix a graph R_n and identification $F_n = \pi_1(R_n)$; a point is then an equivalence class of pairs (g, G), where

- G is a finite connected metric graph with no univalent or bivalent vertices,
- $g\colon R_n \to G$ is a homotopy equivalence.

We normalize the metric so that the sum of the lengths of the edges in G is equal to one; then two pairs (g, G) and (g', G') are *equivalent* if there is an isometry $h\colon G \to G'$ with $h \circ g \simeq g'$. An equivalence class of pairs is called a *marked graph*.

Varying the lengths of edges in a marked graph with k edges and total length one allows one to sweep out an open $(k-1)$-simplex of points in Outer space. Collapsing an edge which is not a loop determines a new open simplex which is a face of the original simplex. We topologize Outer space as the union of these open simplices, modulo these face identifications. A picture of Outer space for $n = 2$ is given in Figure 1.

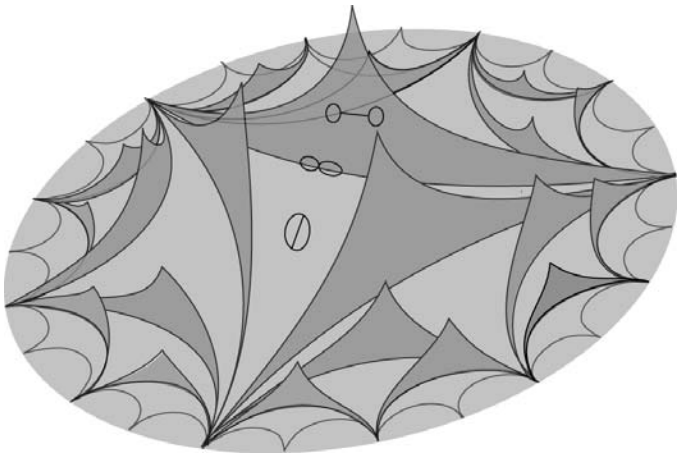


Figure 1. Outer space in rank 2.

2.2. Virtual cohomological dimension. The first cohomological results about automorphism groups of free groups were finiteness results, which followed directly by considering the topology and combinatorial structure of Outer space. Unlike homogeneous spaces and Teichmüller spaces, Outer space and its quotient are not manifolds,

as can be seen already for $n = 2$. However, observed through the prism of cohomology they resemble manifolds in several important ways, including satisfying various finiteness properties and a type of duality between homology and cohomology. As a start, we have

Theorem 2.1 ([12]). *Outer space is contractible of dimension* $3n - 4$*, and* $\mathrm{Out}(F_n)$ *acts with finite stabilizers.*

$\mathrm{Out}(F_n)$ has torsion-free subgroups of finite index, which by the above theorem must act freely on $\mathcal{O}_n$. The quotient of $\mathcal{O}_n$ by such a subgroup Γ is thus an aspherical space with fundamental group Γ, and the following corollary follows immediately:

Corollary 2.2. *The cohomological dimension of any torsion-free subgroup of finite index in* $\mathrm{Out}(F_n)$ *is at most* $3n - 4$.

Serre proved that in fact the cohomological dimension of a torsion-free subgroup of finite index in any group is independent of the choice of subgroup; this is called the *virtual cohomological dimension*, or VCD, of the group. As in the case of $\mathrm{GL}(n, \mathbb{Z})$ and mapping class groups, the quotient of $\mathcal{O}_n$ by $\mathrm{Out}(F_n)$ is not compact, and the dimension of the homogeneous space does not give the best upper bound on the virtual cohomological dimension. A solution to this problem for $\mathrm{Out}(F_n)$ is given by considering an equivariant deformation retract of $\mathcal{O}_n$, called the *spine of Outer space*. This spine can be described as the geometric realization of the partially ordered set of open simplices of $\mathcal{O}_n$, so has one vertex for every homeomorphism type of marked graph with fundamental group F_n and one k-simplex for every sequence of k forest collapses.

Theorem 2.3 ([12]). *The spine of Outer space is an equivariant deformation retract of Outer space. It has dimension* $2n - 3$ *and the quotient is compact.*

This theorem allows one to compute the virtual cohomological dimension of $\mathrm{Out}(F_n)$ precisely:

Corollary 2.4. *The virtual cohomological dimension of* $\mathrm{Out}(F_n)$ *is equal to* $2n - 3$.

Proof. For $i > 1$, let λ_i be the automorphism of $F_n = F\langle x_1, \ldots, x_n\rangle$ which multiplies x_i by x_1 on the left and fixes all other x_j. Similarly, define ρ_i to be the automorphism which multiplies x_i by x_1 on the right. The subgroup of $\mathrm{Out}(F_n)$ generated by the λ_i and ρ_i is a free abelian subgroup of $\mathrm{Out}(F_n)$ of dimension $2n - 3$, giving a lower bound on the virtual cohomological dimension. The upper bound is given by the dimension of the spine. □

2.3. Finite generation of homology. The fact that the quotient of the spine by $\mathrm{Out}(F_n)$ is compact implies immediately that any torsion-free finite-index subgroup is the fundamental group of an acyclic space with only finitely many cells in each dimension, and in particular its homology is finitely generated in all dimensions. This implies the same result for the entire group $\mathrm{Out}(F_n)$:

Corollary 2.5. *The homology of* $\mathrm{Out}(F_n)$ *is finitely-generated in all dimensions.*

The focus of this article is cohomology, but we would like to point out that Outer space and its spine can also be used to prove properties of $\mathrm{Out}(F_n)$ which are not strictly cohomological. As an example, we note that any finite subgroup of $\mathrm{Out}(F_n)$ can be realized as automorphisms of a finite graph with fundamental group F_n (see, e.g. [11]). This is equivalent to saying that any finite subgroup of $\mathrm{Out}(F_n)$ fixes some vertex of the spine of Outer space, so compactness of the quotient immediately gives the following information about the subgroup structure of $\mathrm{Out}(F_n)$.

Corollary 2.6. $\mathrm{Out}(F_n)$ *has only finitely many conjugacy classes of finite subgroups.*

Going even farther afield, we remark that the very concrete description of the spine in terms of graphs and forest collapses allows one to determine the local structure quite precisely. In particular, a neighborhood of a *rose*, i.e. a graph with one vertex and n edges, is easily identified with the space of trees with $2n$ labeled leaves, and can be used as a model for the space of phylogenetic trees in biology. This neighborhood can be given a metric of non-positive curvature, which has a computational advantage for applications to biology (see [3]).

3. The bordification and duality

There is another approach to resolving difficulties arising from the fact that the action of $\mathrm{Out}(F_n)$ on Outer space is not cocompact. Instead of finding a spine inside Outer space, one can extend Outer space and the action of $\mathrm{Out}(F_n)$ by adding cells "at infinity" to produce a larger space whose quotient is compact. This was the approach taken by Borel and Serre in their work on arithmetic groups. They defined a *bordification* of the homogeneous space and used Poincaré–Lefschetz duality for this "manifold with corners" to prove that arithmetic groups satisfy a form of duality between homology and cohomology. Specifically, a group Γ is said to be a *duality group* if there is a module D, integer d and isomorphisms $H^i(\Gamma; M) \to H_{d-i}(\Gamma; M \otimes D)$ for any integer i and Γ-module M. If Γ is a duality group, then the integer d is equal to the virtual cohomological dimension, and this is how Borel and Serre determined the (virtual) cohomological dimension of arithmetic groups.

Although Outer space is not a manifold, Bestvina and Feighn showed that torsion-free finite index subgroups of $\mathrm{Out}(F_n)$ are duality groups, i.e. $\mathrm{Out}(F_n)$ is a *virtual* duality group. They accomplished this by defining a bordification $\widehat{\mathcal{O}}_n$ of Outer space and studying its topology at infinity. This bordification has the structure of a locally finite cell complex on which $\mathrm{Out}(F_n)$ acts with finite stabilizers.

3.1. Cells in the bordification. There are only a finite number of orbit classes of open simplices in $\mathcal{O}_n$, leading one to expect that the quotient should be compact. The reason it is not is that we are leaving out some of the faces of simplices; we go to a

face by collapsing edges in the graph, but we are not allowed to reduce the rank of the graph. One might think of simply adding the missing faces to achieve cocompactness, but this destroys essential features of the action: in particular, the result is not locally finite, and simplex stabilizers are infinite. Bestvina and Feighn found a way around this by keeping track of exactly how subgraphs degenerate as you approach a missing face. Thus there is a cell "at infinity" for each marked metric graph G and sequence of nested subgraphs $G = G_0 \supset G_1 \supset \cdots \supset G_k$. Each subgraph G_i comes with its own metric of volume 1, and G_{i+1} is spanned by the edges of length zero in G_i. The idea is that the sequence consists of subgraphs which are collapsing to zero faster and faster, and the metrics keep track of the direction one is going as one approaches infinity.

3.2. Virtual duality. Bieri and Eckman showed that a group is a virtual duality group if and only if the cohomology of the group with coefficients in its integral group ring vanishes in all but one dimension, where it is free. The cohomology of $\mathrm{Out}(F_n)$ with coefficients in its integral group ring is isomorphic to the cohomology with compact supports of the bordification $\widehat{\mathcal{O}}_n$, and this in turn can be shown to satisfy Bieri and Eckmann's criteria by showing that $\widehat{\mathcal{O}}_n$ is sufficiently connected at infinity. This is what Bestvina and Feighn prove, using Morse theory techniques:

Theorem 3.1 ([2]). *The bordification $\widehat{\mathcal{O}}_n$ of Outer space is $(2n-3)$-connected at infinity.*

Corollary 3.2 ([2]). $\mathrm{Out}(F_n)$ *is a virtual duality group.*

4. The Degree Theorem and rational homology stability

We now turn attention to the group $\mathrm{Aut}(F_n)$. An advantage that $\mathrm{Aut}(F_n)$ has over $\mathrm{Out}(F_n)$ is that it comes equipped with natural inclusions $\mathrm{Aut}(F_n) \to \mathrm{Aut}(F_{n+1})$. The analogous inclusions in the general linear case induce maps $H_k(\mathrm{GL}(n,\mathbb{Z})) \to H_k(\mathrm{GL}(n+1,\mathbb{Z}))$ which were shown by Charney to be isomorphisms for n sufficiently large with respect to k [6]. The fact that the homology of $\mathrm{GL}(n,\mathbb{Z})$ stabilizes in this way serves to considerably simplify the problem of computing the homology. For example, one may determine $H_k(\mathrm{GL}(n,\mathbb{Z}))$ for any large n by performing the computations with relatively small values of n, where the size of the computation is more manageable. A more subtle and more powerful advantage is that one may work instead with the stable groups $\mathrm{GL}_\infty(\mathbb{Z}) = \lim_{n\to\infty} \mathrm{GL}(n,\mathbb{Z})$ which carry additional multiplicative structure and are amenable to homotopy theoretic methods such as the plus construction.

There is a construction completely analogous to the construction of Outer space using basepointed graphs, where the basepoint may be at a vertex or in the interior of an edge. This space $\mathcal{A}_n$ is also contractible [16], has a cocompact spine, and acts as a homogeneous space for $\mathrm{Aut}(F_n)$, which acts with finite stabilizers. (A French

colleague suggested that the space $\mathcal{A}_n$ should be called "Autre espace," but the name "Auter space" seems to have taken hold instead.) An advantage to this space is that the basepoint determines a natural Morse function on a marked metric graph, and we can use parameterized Morse theory methods to study the space. The basepoint also allows us to define a filtration of $\mathcal{A}_n$ by highly connected subspaces which, as we will see, are very useful in homology calculations and in particular for proving homology stability theorems.

4.1. The Degree Theorem. We will filter $\mathcal{A}_n$ by the *degree* of a marked graph, where the degree is defined as the number of vertices away from the basepoint counted with multiplicity (multiplicity is the valence of the vertex minus 2). The degree of a graph with fundamental group F_n is then equal to $2n$ minus the valence of the basepoint. Thus a rose has degree zero, a graph with one trivalent vertex away from the basepoint has degree one, and any graph with basepoint in the interior of an edge has degree $2n-2$. The fact that $\mathrm{Aut}(F_n)$ is generated by Nielsen automorphisms, which can be modeled by a homotopy equivalence which wraps one leaf of a rose around another, implies that the degree 1 subspace of $\mathcal{A}_n$ is connected. An analogous statement is true for higher degrees:

Theorem 4.1 (Degree Theorem, [18]). *The subspace $\mathcal{A}_{n,k}$ of Auter space consisting of marked graphs of degree at most k is $(k-1)$-connected.*

Thus $A_{n,k}$ acts as a kind of k-skeleton for Auter space. The action of $\mathrm{Aut}(F_n)$ on $\mathcal{A}_n$ changes the marking on a graph but not the homeomorphism type, so that it preserves the degree, i.e. restricts to an action on $\mathcal{A}_{n,k}$. Since $\mathcal{A}_{n,k}$ is $(k-1)$-connected and $\mathrm{Aut}(F_n)$ acts with finite stabilizers, the homology of $\mathrm{Aut}(F_n)$ with trivial rational coefficients can be identified with the rational homology of the quotient in dimensions less than k.

Because a degree k graph has only k vertices away from the basepoint (counted with multiplicity), if we ignore loops at the basepoint there are only a finite number of possibilities for such a graph. Thus the map from $\mathcal{A}_{n,k}$ to $\mathcal{A}_{n+1,k}$ given by attaching an extra loop at the basepoint is a homeomorphism for n large. As an immediate consequence, we see that the map $H_{k-1}(\mathrm{Aut}(F_n);\mathbb{Q}) \to H_{k-1}(\mathrm{Aut}(F_{n+1});\mathbb{Q})$ induced by inclusion is an isomorphism on homology n for n large.

For computational purposes, it is obviously advantageous to know exactly how large n has to be, i.e. to have the best possible bound on the stability range. An easy Euler characteristic argument shows that the map from $\mathcal{A}_{n,k}$ to $\mathcal{A}_{n+1,k}$ is a homeomorphism for $n \geq 2k$; for example for any $n \geq 4$, the degree 2 subcomplex of the spine has a contractible quotient consisting of 7 triangles glued together as in Figure 2. As a consequence we can say that $\mathrm{Aut}(F_n)$ satisfies *homology stability with slope* 2. The slope can in fact be improved to $3/2$ fairly easily by showing that the quotients in this range, though not actually homeomorphic, are nevertheless homotopy equivalent [18]; further improvement is possible using some calculations in rank 3. The best known result at this time is the following.

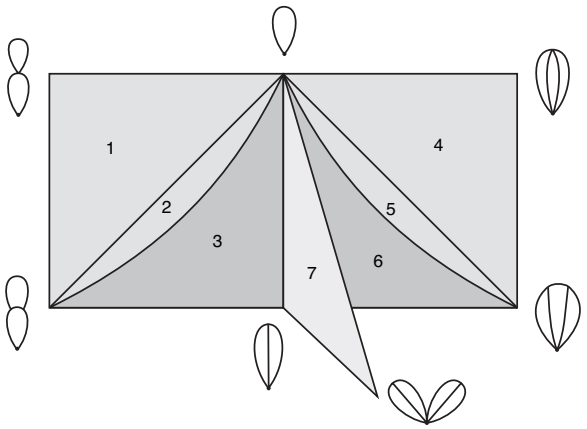


Figure 2. The degree 2 quotient.

Theorem 4.2 ([19]). *The map* $H_{k-1}(\mathrm{Aut}(F_n);\mathbb{Q}) \to H_{k-1}(\mathrm{Aut}(F_{n+1});\mathbb{Q})$ *induced by inclusion is an isomorphism on homology n for* $n \geq 5(k+1)/4$.

Homology stability is a property also shared by surface mapping class [15], with appropriate inclusions. The exact slope, for trivial rational coefficients, is known for both $\mathrm{GL}(n,\mathbb{Z})$ and for mapping class groups of bounded surfaces, but the question remains open for automorphism groups of free groups.

5. Sphere complexes and integral homology stability

5.1. Presentations of $\mathrm{Aut}(F_n)$ and stability for H_1. The first integral homology of $\mathrm{Aut}(F_n)$ is isomorphic to $\mathbb{Z}/2$ for all $n \geq 3$, as can be seen by abelianizing a presentation. There are several different presentations of $\mathrm{Aut}(F_n)$ available, including presentations due to J. Nielsen, B. Neumann, and J. McCool. A new way of obtaining a presentation is supplied by the Degree Theorem, which tells us that the degree 2 subcomplex of Auter space is simply-connected. The method of K. Brown [5] for calculating a presentation from the action of a group on a simply-connected CW-complex can then be used to find a presentation. This was carried out in [1]. The generators, for any $n \geq 4$, are the stabilizers of the seven graphs pictured in Figure 2, with appropriate numbers of loops added at the basepoint. The relations are the relations in these stabilizers, together with relations given by inclusions of edge stabilizers into vertex stabilizers and by inclusions composed with conjugations.

5.2. Quillen's method. More delicate techniques are needed to show that the k-th homology of $\mathrm{Aut}(F_n)$ with trivial integral coefficients stabilizes for n large. The general idea is borrowed from Quillen's work on homology stability for general linear

groups of fields. The simplest possible setup for using this method to prove homology stability for a sequence of groups $\{G_n\}$ is provided by having contractible simplicial complexes X_n and actions of G_n on X_n which are transitive on p-simplices for all p. The stabilizer of a p-simplex should be isomorphic to G_{n-p-1} and the quotient of X_n by the action should be highly connected. Given these conditions, one looks at the equivariant homology spectral sequence for this action, which has

$$E^1_{p,q} = \begin{cases} \bigoplus_{\sigma_p} H_q(\mathrm{stab}(\sigma_p)) & p \geq 0, \\ H_q(G_n) & p = -1, \end{cases}$$

where the direct sum is over all orbits of p-simplices. This spectral sequence converges to 0, and the map $H_k(G_{n-1}) \to H_k(G_n)$ induced by inclusion appears as the map $d^1 \colon E^1_{0,k} \to E^1_{-1,k}$. By induction, we may assume that we understand what happens below the k-th row of the spectral sequence, specifically that the E^2 page vanishes below that row. Since the entire spectral sequence converges to zero, this implies that the d^1 map in question must be onto. Further argument is needed to show that d^1 is injective; this can either be done by increasing the dimension n and applying induction again or by carefully analyzing the next d^1 map, $d^1 \colon E^1_{k,1} \to E^1_{k,0}$ and showing that this is the zero map.

In practice, conditions are usually not quite this nice: the space X_n may not be contractible, the stabilizers may not be precisely G_{n-p-1}, the action may not be transitive on simplices, etc. These difficulties can sometimes be overcome at the cost of introducing further spectral sequence arguments and/or settling for a weaker stability range.

For $G_n = \mathrm{Aut}(F_n)$, homology stability was first proved by Hatcher and Vogtmann in [18] using a complex of free factorizations of the free group F_n. They reproved this theorem in [20] using a different complex, which we describe below. The second paper (together with the erratum [21]) also contains a proof that the map from $\mathrm{Aut}(F_n)$ to $\mathrm{Out}(F_n)$ induces isomorphisms on k-th homology for n large.

The complexes used in [20] involve isotopy classes of 2-spheres embedded in a connected sum of n copies of $S^2 \times S^1$, with a small ball removed. This 3-manifold M_n has fundamental group F_n, and Laudenbach proved that the natural map from the mapping class group of M_n to $\mathrm{Aut}(F_n)$ is surjective, with kernel a 2-torsion subgroup generated by Dehn twists along 2-spheres. This kernel acts trivially on the set of isotopy classes of 2-spheres in M_n, so we obtain an action of $\mathrm{Aut}(F_n)$ on the complex formed by taking a k-simplex for every set of $k+1$ isotopy classes which can be represented by disjoint spheres. The idea of using the complex of isotopy classes of 2-spheres in M_n originated in [16], where Hatcher established many of the basic tools needed for working with such complexes.

A vertex in the complex used in [20] is a non-separating sphere together with an extra, "enveloping" sphere which cuts the sphere and the boundary of M_n off from the rest of the manifold. An alternate description of such a vertex is obtained by using embedded arcs to *tether* each side of the sphere to the boundary of M_n. A set of

$k+1$ isotopy classes of tethered 2-spheres forms a k-simplex if representatives can be found so that all spheres and tethers are disjoint. The stabilizer of a vertex is then isomorphic to $\mathrm{Aut}(F_{n-1})$, and the following theorem allows us to apply Quillen's method:

Theorem 5.1 ([20]). *The complex of isotopy classes of tethered 2-spheres in M_n is $(n-3)/2$-connected.*

As a result, we obtain the following homology stability theorem:

Theorem 5.2 ([20]). *The map $i_*\colon H_k(\mathrm{Aut}(F_n)) \to H_k(\mathrm{Aut}(F_{n+1}))$ induced by the natural inclusion is an isomorphism for $n \geq 2k+2$.*

We are also interested in showing that the map $p_*\colon H_k(\mathrm{Aut}(F_n)) \to H_k(\mathrm{Out}(F_n))$ induced by the natural projection is an isomorphism for n large. In the course of proving this, we are forced to consider the mapping class group of a connected sum of n copies of $S^1 \times S^2$ with $s \geq 0$ balls removed, modulo Dehn twists on 2-spheres. For $s = 0$ this is $\mathrm{Out}(F_n)$. We prove the homology in dimension k is independent of n and s for n and s sufficiently large. In particular, we obtain

Theorem 5.3 ([20], [21]). *The map $p_*\colon H_k(\mathrm{Aut}(F_n)) \to H_k(\mathrm{Out}(F_n))$ is an isomorphism for $n \geq 2k+4$.*

The idea of using "tethers" to tie geometric objects to a basepoint turns out to be useful in other contexts. The extra structure obtained from the tethers has the effect that the conditions for applying Quillen's method are close to ideal, so that the spectral sequence arguments needed to prove homology stability are relatively simple. In particular, tethers have led to simplified proofs of homology stability for mapping class groups of surfaces and braid groups, as well as to new proofs that several related series of groups have homology stability [17].

5.3. Galatius' theorem. Since we know that the homology of $\mathrm{Aut}(F_n)$ stabilizes, the next problem is then to compute the stable homology. Computations in dimensions less than 7 were done in [19], and produced no stable rational homology classes. Igusa showed that the map from the stable rational homology of $\mathrm{Aut}(F_n)$ to that of $\mathrm{GL}(n,\mathbb{Z})$ is the zero map [22]. This evidence led to the conjecture that the stable rational homology is trivial. On the other hand, Hatcher showed that the stable homology contains the stable homology of the symmetric group Σ_n as a direct factor, so there are lots of torsion classes [16]. The entire situation has recently been resolved by S. Galatius [13] using methods adapted from Madsen and Weiss' work on the stable homology of mapping class groups.

The commutator subgroup of $\mathrm{Aut}(F_\infty)$ is a perfect normal subgroup, so that Quillen's plus construction can be applied to the classifying space $B\,\mathrm{Aut}(F_\infty)$. The resulting space $B\,\mathrm{Aut}(F_\infty)^+$ is an infinite loop space whose homology is equal to the stable homology of $\mathrm{Aut}(F_n)$. The natural inclusions of the symmetric groups Σ_n

into $\mathrm{Aut}(F_n)$ induce an infinite loop space map $B\Sigma_\infty^+ \to B\,\mathrm{Aut}_\infty^+$, and a theorem of Barratt–Priddy and Quillen says that $B\Sigma_\infty^+$ is homotopy equivalent to $\Omega^\infty S^\infty$. The space $\Omega^\infty S^\infty$ is the most fundamental example of an infinite loop space; its homotopy groups are the stable homotopy groups of spheres. In [13] Galatius proves that $B\,\mathrm{Aut}(F_\infty)^+$ is also homotopy equivalent to $\Omega^\infty S^\infty$, showing in particular that the symmetric group and the automorphism group of a free group have the same stable homology. The proof relies on the contractibility of Outer space and the homology stability results of [20] and [21]. Galatius proceeds by defining maps of $B\,\mathrm{Out}(F_n)$ to a certain "graph spectrum" E, whose n-th space is the space of all graphs in $\mathbb{R}^n$. He proves that after passing to infinite loop spaces this map becomes a homotopy equivalence, and then that $\Omega^\infty E$ is in fact homotopy equivalent to $\Omega^\infty S^\infty$. The homology of $\Omega^\infty S^\infty$ is torsion, so that the stable rational homology of $\mathrm{Aut}(F_n)$ is trivial as conjectured.

6. Graph complexes and unstable homology

Though the stable homology of $\mathrm{Aut}(F_n)$ and $\mathrm{Out}(F_n)$ has been completely determined, at this writing the unstable homology is still largely mysterious. In this section we consider the unstable *rational* homology.

6.1. Low-dimensional calculations. The simplices in the spines of Outer space and Auter space naturally group themselves into cubes, giving these spines the structure of cube complexes. Specifically, an m-dimensional cube corresponds to a marked graph (g, G) together with a subforest Φ of G with m edges, since the set of simplices which can be obtained by collapsing the edges in the subforest in any one of the 2^{m-1} possible orders fit together to form a cube.

The quotient of a cube by a linear map is a rational homology cell, so that the cube complex structure on the spine descends to a "cell structure" on the quotient, which can be used to compute the rational homology of $\mathrm{Out}(F_n)$ and $\mathrm{Aut}(F_n)$. For $\mathrm{Aut}(F_n)$, the Degree Theorem can be used to reduce the number of cubes one must consider when computing the kth homology, and further reductions are possible by examining the structure of the quotient. In the end it is possible for $k = 2, 3$ and even 4 to do the computations by hand. For $k > 4$, however, the aid of a computer becomes essential. Computations for both $\mathrm{Aut}(F_n)$ and $\mathrm{Out}(F_n)$ for $k \leq 7$ were carried out by Hatcher and Vogtmann, Jensen and Gerlits. They showed that that $H_k(\mathrm{Aut}(F_n); \mathbb{Q}) = H_k(\mathrm{Out}(F_n); \mathbb{Q}) = 0$ for $k \leq 7$ except that $H_4(\mathrm{Aut}(F_4); \mathbb{Q}) = H_4(\mathrm{Out}(F_4); \mathbb{Q}) \cong \mathbb{Q}$ [19]. This agrees with Galatius' theorem in the stable range, and gives a tantalizing glimpse into the unstable homology. We now understand this very interesting non-trivial unstable homology class in a much more general context, as we will see below.

6.2. Graph homology of a cyclic operad. In [23], [24] Kontsevich found a remarkable correspondence between the cohomology of certain infinite-dimensional

symplectic Lie algebras and the homology of outer automorphism groups of free groups. This Lie algebra cohomology can be computed using the subcomplex of the Chevalley–Eilenberg complex spanned by symplectic invariants. The connection with $\mathrm{Out}(F_n)$ is made via Weyl's invariant theory, which allows one to interpret the complex of symplectic invariants as a chain complex indexed by finite graphs, where the vertices of these graphs are decorated by elements of the Lie operad. The homology of this graph complex can then be interpreted in terms of Outer space; this is carried out explicitly in [8].

The same formalism using the associative operad in place of the Lie operad gives a chain complex of "ribbon graphs", which computes the homology of surface mapping class groups. The commutative operad gives rise to a type of graph homology which includes information about diffeomorphism groups of odd-dimensional homology spheres. Kontsevich's construction in fact makes sense using any cyclic operad to decorate the vertices of graphs (see [8]), and it would be interesting to study the functorial properties of the resulting homology theories.

The connection with Lie algebra homology reveals new structure on the level of chain complexes. Specifically, the graph homology chain complex for any cyclic operad supports a Lie bracket and cobracket, which were studied in [10]; the Lie bracket can be shown to correspond to the classical Schouten bracket on the Lie algebra. The bracket and cobracket do not in general form a compatible Lie bialgebra structure, but do on the subcomplex spanned by connected graphs with no separating edges. For the associative and Lie operads, this subcomplex is quasi-isomorphic to the whole complex, so in particular has the same homology. For the commutative operad, this is not true, but the Lie bracket and cobracket do induce a bracket and cobracket on an appropriate quotient complex which is quasi-isomorphic to the whole complex [8]. These brackets come from a second boundary operator on the graph complex, and measure the deviation of this boundary operator from being a derivation (resp. coderivation). This second boundary operator anti-commutes with the standard boundary operator, so induces a map on graph homology. The Lie bracket and cobracket vanish on the level of homology, making graph homology together with this induced map into a differential graded algebra.

6.3. Morita cycles. Kontsevich's work also led to new discoveries by S. Morita, who had been studying some of the same Lie algebras in his work on surface mapping class groups. In particular, Morita found an infinite sequence of cocycles for these Lie algebras based on his "trace" map and showed that the first of these cocycles is non-trivial on cohomology [28]. Via Kontsevich's theorem, this cocycle produces a nontrivial homology class in $H_4(\mathrm{Out}(F_4); \mathbb{Q})$. Since we know the rational homology $H_4(\mathrm{Out}(F_4); \mathbb{Q})$ is one-dimensional, this class in fact completely computes the homology in this dimension.

Conant and Vogtmann translated Morita's cocycles into cocycles on the complex of Lie graphs. They then showed that the combinatorial information contained in an oriented graph with vertices decorated by basic elements of the Lie operad is captured

more simply by a trivalent graph together with a subforest whose edges are ordered. This allows one to reinterpret the Morita cocycles directly in terms of such forested graphs, and to give a quick proof that the first of Morita's cocycles is non-trivial. With a little more work it can be used to show that second one is also non-trivial, giving a rational homology class in $H_8(\mathrm{Out}(F_6);\mathbb{Q})$ [9]. Morita reports that R. Ohashi has recently shown that in fact $H_8(\mathrm{Out}(F_6);\mathbb{Q}) \cong \mathbb{Q}$, so that this class gives all of the homology [27].

There is a Morita cocycle corresponding to every graph consisting of two vertices joined by an odd number of edges. Both Morita and Conant-Vogtmann found generalizations which give a class in $H_{2k+r-2}(\mathrm{Out}(F_{k+r}))$ for every odd-valent graph of rank r with k vertices. One expects that all of these should be non-trivial classes, and it is possible that they determine all of the unstable rational homology of $\mathrm{Out}(F_n)$. We would therefore like to understand these cocycles as well as possible. Since they determine rational homology classes in the homology of $\mathrm{Out}(F_n)$, we should be able to find representatives in the quotient of the spine of Outer space and in the quotient of the bordification of Outer space, both of which are rationally acyclic spaces for $\mathrm{Out}(F_n)$. This is indeed possible, as is shown in [7]. In the case of the bordification, the cycles are found "at infinity." Recall that a cell at infinity is given by a marked filtered graph (see section 3). The action of $\mathrm{Out}(F_n)$ is transitive on markings, so that they disappear in the quotient. The first Morita cycle is the quotient of the single cell shown in Figure 3.

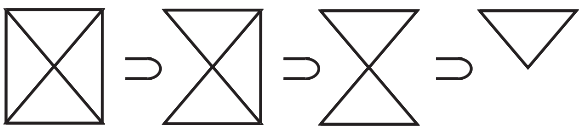


Figure 3. Generator of $H_4(\mathrm{Out}(F_4);\mathbb{Q})$ in the bordification.

For the description of this class in terms of the spine, recall that the spine of Outer space has the structure of a cube complex, where a cube is given by a marked graph together with a subforest. Again, the action of $\mathrm{Out}(F_n)$ is transitive on the markings, and the first Morita cycle is the union of the quotient of the three cubes shown in Figure 4.

Conant and Vogtmann use these descriptions in [7] to show that all of the Morita classes are unstable in the strongest possible sense: they vanish when the rank of the free group increases by one.

6.4. Rational Euler characteristic. A homological invariant of infinite groups G which is often easier to compute than the complete cohomology is the rational Euler characteristic $\chi(G)$. This is defined as the usual alternating sum of the Betti numbers for any torsion-free subgroup of finite index, divided by the index of the subgroup, and was shown by Serre to be independent of the choice of the subgroup.

Figure 4. Generator of $H_4(\mathrm{Out}(F_4);\mathbb{Q})$ in the spine.

For $\mathrm{GL}(n,\mathbb{Z})$ Harder showed that the rational Euler characteristic vanishes for all $n \geq 3$, and in general the rational Euler characteristics of arithmetic groups are closely related to values of zeta functions. For mapping class groups of closed surfaces the rational Euler characteristic vanishes for surfaces of odd genus, but for even genus it alternates in sign and is basically given by the classical Bernoulli numbers, as was shown by Harer and Zagier in [14].

The rational Euler characteristic of a group can be computed by finding a contractible complex on which the group acts cocompactly with finite stabilizers, and calculating the alternating sum, over all orbits of cells σ, of the terms

$$\frac{(-1)^{\dim(\sigma)}}{|\,\mathrm{stab}(\sigma)|}.$$

This was done for $\mathrm{Out}(F_n)$ in [29] using the spine of outer space. The result was a generating function for $\chi(\mathrm{Out}(F_n))$ built from standard generating functions for counting graphs and forests in graphs. Using this generating function, $\chi(\mathrm{Out}(F_n))$ was computed explicitly for values of n up to 100. It is strictly negative in all cases computed and seems to grow in absolute value faster than exponentially. Smillie and Vogtmann proved that $\chi(\mathrm{Out}(F_n))$ is non-zero for all even n, and computed the p-power of the denominator for many primes p.

A different approach to the problem of counting graphs and forests was given by Kontsevich, who produced an integral formula for the rational Euler characteristic of $\mathrm{Out}(F_n)$ using techniques of perturbative series and Feynman diagrams. He also produced integral formulas for the rational Euler characteristic of mapping class groups which recapture the relation with Bernoulli numbers found by Harer and Zagier. In the case of $\mathrm{Out}(F_n)$, neither the generating function nor the integral formula make it clear what the asymptotic growth rate of $\chi(\mathrm{Out}(F_n))$ might be, or even whether it is non-zero for all n. A non-zero, quickly growing Euler characteristic would indicate the presence of a large amount of unstable homology.

7. IA automorphisms and the IA quotient of Outer space

At the beginning of this article we noted the existence of a natural map from $\mathrm{Out}(F_n)$ onto $\mathrm{GL}(n,\mathbb{Z})$. The kernel of this map is called the subgroup of *IA automorphisms*

because it consists of automorphisms which induce the *identity* on the *abelianization* of F_n. It is clearly a natural object to study if one is trying to understand the relation between $\mathrm{Out}(F_n)$ and $\mathrm{GL}(n,\mathbb{Z})$. Magnus found a finite generating set for IA_n in 1934, and asked at the same time whether IA_n was finitely presentable [26].

There is an interesting application of the non-vanishing of the rational Euler characteristic of $\mathrm{Out}(F_n)$ to the study of IA_n. If the homology of IA_n were finitely generated then the rational Euler characteristic of IA_n would be defined, and the short exact sequence

$$1 \to IA_n \to \mathrm{Out}(F_n) \to \mathrm{GL}(n,\mathbb{Z}) \to 1$$

would result in the equation $\chi(IA_n)\chi(\mathrm{Out}(F_n)) = \chi(\mathrm{GL}(n,\mathbb{Z}))$. However, we know that $\chi(\mathrm{GL}(n,\mathbb{Z})) = 0$ for $n \geq 3$, while $\chi(\mathrm{Out}(F_n))$ is non-zero, at least for n even. Thus we can conclude that the homology of IA_n is not finitely generated in some dimension.

If IA_n was finitely presentable, that would imply that the second homology is finitely generated. The argument in the previous paragraph shows that the homology is not finitely generated in *some* dimension, but gives no definite conclusion about dimension 2. McCool and Krstic finally answered Magnus' question for $n = 3$ in 1997, by showing that IA_3 is *not* finitely presentable [25]. Recently Bestvina, Bux and Margalit have shown that the top-dimensional homology of IA_n vanishes, while the codimension one homology is infinitely generated. They prove this by using Morse theory to study the topology of the quotient of Outer space by IA_n, which is an aspherical space with fundamental group IA_n. This work implies the McCool–Krstic result, since it says that $H_3(IA_3) = 0$ and $H_2(IA_3)$ is infinitely generated, but still leaves open the question of finite presentability for $n \geq 4$.

8. Further reading

In this article I have focused on cohomological properties of automorphism groups of free groups, but there are many other areas in which our knowledge of these groups is rapidly expanding. These include, for example, the subgroup structure, metric theory and rigidity properties. I wrote two other survey articles which address some of these advances. The first paper [30] gives a more detailed introduction to Outer space and related spaces and mentions other powerful techniques such as Bestvina–Handel's train tracks, as well as many applications to the study of $\mathrm{Out}(F_n)$ and $\mathrm{Aut}(F_n)$. It also contains a fairly extensive bibliography and more thorough references for work on automorphisms of free groups. The focus of the more recent paper [4], which is joint with Martin Bridson, is a discussion of open problems in the field.

References

[1] Armstrong, Heather, Forrest, Brad, and Vogtmann, Karen, A presentation for $\mathrm{Aut}(F_n)$. Preprint, 2006.

[2] Bestvina, Mladen, and Feighn, Mark, The topology at infinity of $\mathrm{Out}(F_n)$. *Invent. Math.* **140** (3) (2000), 651–692.

[3] Billera, Louis J., Holmes, Susan P., and Vogtmann, Karen, Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** (4) (2001), 733–767.

[4] Bridson, Martin R., and Vogtmann, Karen, Automorphisms of free groups, surface groups and free abelian groups. Preprint, 2005; math.GR/0507602.

[5] Brown, Kenneth S., Presentations for groups acting on simply-connected complexes. *J. Pure Appl. Algebra* **32** (1) (1984), 1–10.

[6] Charney, Ruth M., Homology stability of GL_n of a Dedekind domain. *Bull. Amer. Math. Soc.* (*N.S.*) **1** (2) (1979), 428–43.

[7] Conant, James, and Vogtmann, Karen, The Morita classes are stably trivial. Preprint, 2006.

[8] Conant, James, and Vogtmann, Karen, On a theorem of Kontsevich. *Algebr. Geom. Topol.* **3** (2003), 1167–1224.

[9] Conant, James, and Vogtmann, Karen, Morita classes in the homology of automorphism groups of free groups. *Geom. Topol.* **8** (2004), 1471–1499.

[10] Conant, Jim, and Vogtmann, Karen, Infinitesimal operations on complexes of graphs. *Math. Ann.* **327** (3) (2003), 545–573.

[11] Culler, Marc, Finite groups of outer automorphisms of a free group. In *Contributions to group theory*, Contemp. Math. 33, Amer. Math. Soc., Providence, RI, 1984, 197–207.

[12] Culler, Marc, and Vogtmann, Karen, Moduli of graphs and automorphisms of free groups. *Invent. Math.* **84** (1) (1986), 91–119.

[13] Galatius, Soren, in preparation.

[14] Harer, J., and Zagier, D., The Euler characteristic of the moduli space of curves. *Invent. Math.* **85** (3) (1986), 457–485.

[15] Harer, John L., Stability of the homology of the moduli spaces of Riemann surfaces with spin structure. *Math. Ann.* **287** (2) (1990), 323–334.

[16] Hatcher, Allen, Homological stability for automorphism groups of free groups. *Comment. Math. Helv.* **70** (1) (1995), 39–62.

[17] Hatcher, Allen, and Vogtmann, Karen, Tethers and homology stability. Preprint, 2006.

[18] Hatcher, Allen, and Vogtmann, Karen, Cerf theory for graphs. *J. London Math. Soc.* (2) **58** (3) (1998), 633–655.

[19] Hatcher, Allen, and Vogtmann, Karen, Rational homology of $\mathrm{Aut}(F_n)$. *Math. Res. Lett.* **5** (6) (1998), 759–780.

[20] Hatcher, Allen, and Vogtmann, Karen, Homology stability for outer automorphism groups of free groups. *Algebr. Geom. Topol.* **4** (2004), 1253–1272.

[21] Hatcher, Allen, Vogtmann, Karen, and Wahl, Nathalie, Erratum to: Homology stability for outer automorphism groups of free groups. *Algebr. Geom. Topol.* **6** (2006), 573–579.

[22] Igusa, Kiyoshi, *Higher Franz-Reidemeister torsion*. AMS/IP Stud. Advanced Math. 31, Amer. Math. Soc., Providence, RI, 2002.

[23] Kontsevich, Maxim, Formal (non)commutative symplectic geometry. In *The Gel'fand Mathematical Seminars, 1990–1992*, Birkhäuser Boston, Boston, MA, 1993, 173–187.

[24] Kontsevich, Maxim, Feynman diagrams and low-dimensional topology. In *First European Congress of Mathematics* (Paris, 1992), Vol. II, Progr. Math. 120, Birkhäuser, Basel 1994, 97–121.

[25] Krstić, Sava, and McCool, James, The non-finite presentability of $\mathrm{IA}(F_3)$ and $\mathrm{GL}_2(\mathbf{Z}[t, t^{-1}])$. *Invent. Math.* **129** (3) (1997), 595–606.

[26] Magnus, Wilhelm, Über n-dimensionale Gittertransformationen. *Acta Math.* **64** (1934), 353–367.

[27] Morita, Shigeyuki, Cohomological structure of the mapping class group and beyond. Preprint, 2005; math.GT/0507308.

[28] Morita, Shigeyuki, Structure of the mapping class groups of surfaces: a survey and a prospect. In *Proceedings of the Kirbyfest* (Berkeley, CA, 1998), Geom. Topol. Monogr. 2, Geom. Topol. Publ., Coventry, 1999, 349–406.

[29] Smillie, John, and Vogtmann, Karen, A generating function for the Euler characteristic of $\mathrm{Out}(F_n)$. *J. Pure Appl. Algebra* **44** (1–3) (1987), 329–348.

[30] Vogtmann, Karen, Automorphisms of free groups and outer space. *Geom. Dedicata* **94** (2002), 1–31.

Department of Mathematics, Cornell University, Ithaca, NY 14853-4201, U.S.A.

E-mail: vogtmann@math.cornell.edu

Noncommutative counterparts of the Springer resolution

Roman Bezrukavnikov*

Abstract. Springer resolution of the set of nilpotent elements in a semisimple Lie algebra plays a central role in geometric representation theory. A new structure on this variety has arisen in several representation theoretic constructions, such as the (local) geometric Langlands duality and modular representation theory. It is also related to some algebro-geometric problems, such as the derived equivalence conjecture and description of T. Bridgeland's space of stability conditions. The structure can be described as a noncommutative counterpart of the resolution, or as a t-structure on the derived category of the resolution. The intriguing fact that the same t-structure appears in these seemingly disparate subjects has strong technical consequences for modular representation theory.

Mathematics Subject Classification (2000). Primary 17B50, 18F99, 20G05; Secondary 20G42, 22E67.

Keywords. Derived categories of sheaves and modules, modular Lie algebras, local geometric Langlands duality.

1. Introduction

Springer resolution of the variety of nilpotent elements in a semi-simple Lie algebra is ubiquitous in geometric representation theory. In this article we show that, besides this well-known resolution of singularities, the variety of nilpotents, as well as some other closely related varieties, admits a particular *noncommutative resolution of singularities*, which arises in different representation theoretic and algebro-geometric constructions. Here by a noncommutative resolution of a singular variety Y we mean, following, e.g., [11], a coherent sheaf of associative $\mathcal{O}_Y$ algebras satisfying certain natural conditions, and defined up to a Morita equivalence.

The constructions are related to such subjects as: the (local) geometric Langlands duality program and categorification of representation theory of affine Hecke algebras, representation theory of modular Lie algebras and quantum enveloping algebras at roots of unity, Bridgeland's theory of stability conditions on triangulated categories, and categorical McKay correspondence and generalizations.

Let G be a semi-simple adjoint algebraic group, $\mathfrak{g}$ be its Lie algebra and $\mathcal{N} \subset \mathfrak{g}$ be the variety of nilpotent elements. Let $\mathcal{B}$ be the variety of Borel subalgebras in $\mathfrak{g}$, also known as the flag variety of G, and $\widetilde{\mathcal{N}} = T^*(\mathcal{B})$ be the cotangent bundle to $\mathcal{B}$. The Springer resolution is the moment map $\pi : \widetilde{\mathcal{N}} \to \mathcal{N}$.

*The author is partially supported by NSF grant DMS-0505466 and Sloan Foundation grant 0965-300-L216.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Our noncommutative resolution A of $\mathcal{N}$ comes with an equivalence between the derived category $D(A)$ of modules over A and the derived category $D(\widetilde{\mathcal{N}})$ of coherent sheaves on $\widetilde{\mathcal{N}}$. Thus A is determined uniquely up to Morita equivalence by the t-structure on $D(\widetilde{\mathcal{N}})$ induced by the equivalence, i.e., by the image of the subcategory of A modules in $D(A)$ under the equivalence. We will call this t-structure the *exotic t-structure* and objects of its heart *exotic sheaves*. Thus an exotic sheaf is a complex of coherent sheaves on $\widetilde{\mathcal{N}}$ which corresponds to an A-module under the equivalence $D(A) \cong D(\widetilde{\mathcal{N}})$.

Closely related data first appeared in [2], which can be considered as a contribution to a local version of the geometric Langlands duality program [8], [37], [34]. A typical result of geometric Langlands duality is an equivalence between some derived category of constructible sheaves on a variety related to ${}^LG$ bundles on a curve C and derived category of coherent sheaves on a variety related to G local systems on C; here G and ${}^LG$ are reductive groups, which are dual in the sense of Langlands. In the local version of the theory the curve C is a punctured formal disc $\mathbb{D}$. The role of the moduli stack of ${}^LG$ bundles is played by a homogeneous space for the group ${}^LG$, where ${}^LG((t))$ stands for the group of maps from $\mathbb{D}$ to ${}^LG$ (also known as the formal loop group). An example of such a homogeneous space is the *affine flag variety* $\mathcal{F}\ell$ of ${}^LG$. For an appropriate choice of the category of constructible sheaves, the variety related to G local systems turns out to be $\widetilde{\mathcal{N}}$, or rather the quotient stack $\widetilde{\mathcal{N}}/G$ of $\widetilde{\mathcal{N}}$ by the natural action of G. An equivalence between the derived category of G-equivariant coherent sheaves on $\widetilde{\mathcal{N}}$ and a certain triangulated category of constructible sheaves on $\mathcal{F}\ell$ is proved in [2]. The image of the subcategory of perverse sheaves on $\mathcal{F}\ell$ under this equivalence turns out to consist of *equivariant exotic sheaves*, which are closely related to exotic sheaves (see Section 2.2 below).

Another construction leading to exotic sheaves is related to modular representation theory.

In the second half of the 20th century various geometric methods for representation theory of semi-simple Lie algebras over characteristic zero fields have been developed. One of the culminating points is the Localization Theorem [5], [30], motivated by a conjecture by Kazhdan and Lusztig, which provides an equivalence between the category of modules over a semi-simple Lie algebra $\mathfrak{g}$ with a fixed (integral regular) central character and the category of D-modules on the flag variety $\mathcal{B}$. In the paper [19], motivated by Lusztig's extension [45] of Kazhdan–Lusztig conjectures to the modular setting, we provide a similar result for semi-simple Lie algebras over algebraically closed fields of positive characteristic. More precisely, we establish a derived localization theorem, which is an equivalence between the derived category of appropriately defined D-modules (called crystalline, or PD D-modules) on a flag variety and the derived category of Lie algebra modules, where a part of the center, the so-called Harish-Chandra center, acts by a fixed character.

Furthermore, in the case of positive characteristic there is a close relationship between crystalline D-modules on a smooth variety X and coherent sheaves on the cotangent space T^*X [19], [48]. The algebra of crystalline differential operators has a

huge center provided by invariant polynomials of the p-curvature of a D-module. This allows one to view the differential operators as a sheaf of algebras on the cotangent bundle. This algebra turns out to be an Azumaya algebra. In the case of the flag variety this Azumaya algebra splits on the formal neighborhood of each Springer fiber. Thus the derived localization theorem yields a full embedding from the category of finite dimensional $\mathfrak{g}$-modules with a fixed (integral regular) action of the Harish-Chandra center into the derived category of coherent sheaves on $\widetilde{\mathcal{N}}$. It turns out that the image of this embedding consists precisely of exotic sheaves with proper support. A similar relation is expected between exotic sheaves over a field of characteristic zero and representations of the quantum Kac–De Concini enveloping algebra at a root of unity [32], and also with some class of $\widehat{{}^L\mathfrak{g}}$ modules at the critical level (cf. [4] and [35] respectively); here $\widehat{{}^L\mathfrak{g}}$ stands for the affine Kac-Moody algebra corresponding to the Langlands dual algebra ${}^L\mathfrak{g}$.

Thus exotic sheaves are related, on the one hand, to perverse sheaves on the affine flag variety for the dual group, and on the other hand, to modular Lie algebra representations. Comparison of these two connections allows one to apply the known deep results about weights of Frobenius acting on Ext's between irreducible perverse sheaves to numerical questions about modular representations, thereby providing a strategy for a proof of Lusztig's conjectures from [45]. The conjectures relate the classes of irreducible $\mathfrak{g}$-modules to elements of the *canonical basis* in the Borel–Moore homology of a Springer fiber; thus our work provides a categorification of the canonical bases in (co)homology of Springer fibers. See also Remark 2.21 for an application to representations of quantum groups.

I also would like to point out some parallels between exotic sheaves and objects arising in the work of algebraic geometers studying derived categories of coherent sheaves on algebraic varieties. Exotic sheaves can be described in terms of a certain action of the affine braid group B_{aff} of ${}^LG$ on $D(\widetilde{\mathcal{N}})$. This description can be reformulated in terms of a map from the set of alcoves (connected components of the complement to affine coroot hyperplanes in the dual space to the Cartan algebra of $\mathfrak{g}$ over $\mathbb{R}$) to the set of t-structures on $D(\widetilde{\mathcal{N}})$. Similar data have been used by Bridgeland in [25]–[27] to construct a component in the space of stability conditions [24], on the derived categories of coherent sheaves on certain varieties. See also Examples 2.8, 2.9 below.

The appearance of the affine braid group, which can be interpreted as the fundamental group of the set of regular semisimple conjugacy classes in the dual group ${}^LG(\mathbb{C})$, suggests a possibility that the structures described above admit a natural interpretation via homological mirror duality, which would identify our derived category of coherent sheaves with a certain Fukaya type category, where the action of the affine braid group arises from monodromy of some family over the space of regular semisimple conjugacy classes in ${}^LG(\mathbb{C})$.

Another connection to algebraic geometry is provided by [17] and [41]. As has been noted above, the derived localization theorem can be interpreted as a construction of a noncommutative resolution of the nilpotent cone $\mathcal{N}$ using crystalline differential

operators in positive characteristic. It turns out that for more general resolutions of singularities, which carry an algebraic symplectic form, a non-commutative resolution can be constructed by a similar procedure. The construction involves quantizing the algebraic symplectic variety in characteristic p, and relating modules over the quantization to coherent sheaves. It has been carried out in [17] for crepant resolutions of quotients V/Γ, where V is a vector space equipped with a symplectic form, and Γ is a finite subgroup in $\mathrm{Sp}(V)$; this yields a particular case of the so-called *categorical McKay correspondence*. The particular case when Γ is the symmetric group on n letters acting on $V = (\mathbb{A}^2)^n$ is related to representations of the rational Cherednik algebra [16]. In Kaledin's work [41] the construction is generalized to more general symplectic resolutions of singularities.

In the remainder of the text we explain some of these contexts (in the order which is roughly inverse to the above) in some detail.

This text is a mixture of an exposition of published results and announcement of yet unpublished ones; statements for which no reference is provided, and which are not well-known, are to appear in a future publication.

Notations and conventions. Throughout the text we work over an algebraically closed field $\boldsymbol{k}$; when a semi-simple group G is involved, we assume that the characteristic of $\boldsymbol{k}$ is zero or exceeds the Coxeter number of G.

For an algebraic variety X we let $\mathcal{O}_X$ denote the structure sheaf, and $D(X) = D^b(\mathrm{Coh}_X)$ be the bounded derived category of coherent sheaves on X. Given an action of an algebraic group H on X we write $\mathrm{Coh}^H(X)$ for the category of H-equivariant coherent sheaves; given a coherent sheaf of associative $\mathcal{O}_X$ algebras we let $\mathrm{Coh}(X, \mathcal{A})$ be the category of sheaves of coherent $\mathcal{A}$ modules; if $\mathcal{A}$ is H-equivariant for an algebraic group H acting on X, we let $\mathrm{Coh}^H(X, \mathcal{A})$ be the category of H-equivariant sheaves of coherent $\mathcal{A}$-modules. We write $D(X)$, $D^H(X)$, $D(\mathcal{A})$, $D^H(\mathcal{A})$ for the bounded derived category of $\mathrm{Coh}(X)$, $\mathrm{Coh}^H(X)$, $\mathrm{Coh}(X, \mathcal{A})$, $\mathrm{Coh}^H(X, \mathcal{A})$ respectively, and $K(X)$, $K^H(X)$, $K(\mathcal{A})$, $K^H(\mathcal{A})$ for the corresponding Grothendieck groups. In particular, there notations apply for an algebra A finite over the center of finite type.

The functors of pull-back, push-forward etc. between categories of sheaves are understood to be the derived functors.

Acknowledgements. I thank my coauthors S. Arkhipov, V. Ginzburg, D. Kaledin, I. Mirković, V. Ostrik and D. Rumynin for their contribution to the joint results, without which they would have never been accomplished. The project described here was conceived during IAS Special Year in Representation Theory (98/99) led by G. Lusztig. Most of the results have been obtained by unraveling the formulas in Lusztig's papers, thus they owe their existence to him. I have learned Lusztig's results and many other things from M. Finkelberg. I have also benefitted a lot from ideas of I. Mirković and his generosity in sharing them. Conversation with many people were very helpful, the incomplete list includes A. Beilinson, V. Drinfeld, D. Gaitsgory, V. Ginzburg, V. Ostrik. I am very grateful to all these people. Finally, I thank

M. Finkelberg, J. Humphreys, D. Kazhdan and I. Mirković for reading a preliminary version of this text and making helpful comments and suggestions.

2. Noncommutative resolutions and braid group actions

2.1. Braid group actions and noncommutative Springer resolution. Though the motivation for the study of our main object comes from applications to representation theory, we first describe it in the language of algebraic geometry. We briefly recall some ideas from [22], [23], [11].

Let Z be a singular algebraic variety. We refer, e.g., to [11] for the notion of a *crepant* resolution; it is easy to see that resolutions $\pi, \widetilde{\pi}$ described above are crepant.

By a *noncommutative resolution* [11][1] one means a coherent torsion free sheaf A of associative $\mathcal{O}_Z$ algebras, which is generically a sheaf of matrix algebras and has finite homological dimension. There exists also a notion of a noncommutative crepant resolution, see [11]. It has been conjectured in *loc. cit.* that any two crepant resolutions, commutative or not, are derived equivalent, in particular, for any crepant resolution $X \to Z$ and any noncommutative resolution A of Y we have an equivalence $D(X) \cong D(A)$.

2.1.1. The set-up. The notations G, $\mathfrak{g}$, $\mathcal{B}$, $\pi\colon \widetilde{\mathcal{N}} \to \mathcal{N}$ have been defined in the Introduction. Recall that $\widetilde{\mathcal{N}} = T^*(\mathcal{B})$ parametrizes pairs $(\mathfrak{b}, x)$, where $\mathfrak{b} \in \mathcal{B}$ is a Borel subalgebra, and x is the element in the nilpotent radical of $\mathfrak{b}$. The Springer map $\pi\colon \widetilde{\mathcal{N}} \to \mathcal{N}$ is given by $\pi\colon (\mathfrak{b}, x) \mapsto x$. It is embedded in the *Grothendieck simultaneous resolution* $\widetilde{\pi}\colon \widetilde{\mathfrak{g}} \to \mathfrak{g}$, where $\widetilde{\mathfrak{g}}$ is the variety of pairs $(\mathfrak{b}, x)$, $\mathfrak{b} \in \mathcal{B}$, $x \in \mathfrak{b}$, and $\widetilde{\pi}\colon (\mathfrak{b}, x) \mapsto x$. The variety $\widetilde{\mathfrak{g}}$ is smooth, and the map $\widetilde{\pi}$ is proper and generically finite of degree $|W|$, where W is the Weyl group. It factors as the composition of a resolution of singularities $\widetilde{\pi}'\colon \widetilde{\mathfrak{g}} \to \mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$ and the finite projection $\mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h} \to \mathfrak{g}$; here $\mathfrak{h}$ is the Cartan algebra of $\mathfrak{g}$. Let $\mathfrak{g}^{\mathrm{reg}} \subset \mathfrak{g}$ denote the subspace of regular (not necessarily semi-simple) elements, and $\widetilde{\mathfrak{g}}^{\mathrm{reg}}$ be the preimage of $\mathfrak{g}^{\mathrm{reg}}$ in $\widetilde{\mathfrak{g}}$; then $\widetilde{\pi}'$ induces an isomorphism $\widetilde{\mathfrak{g}}^{\mathrm{reg}} \cong \mathfrak{g}^{\mathrm{reg}} \times_{\mathfrak{h}/W} \mathfrak{h}$.

Much of the representation theory of G or $\mathfrak{g}$ is in one way or another related to the geometry of these spaces and maps.

2.1.2. Affine braid group action. For a characterization of our noncommutative resolution we need to introduce some more notation.

Let Λ be the root lattice of G. For $\lambda \in \Lambda$ we will write $\mathcal{O}(\lambda)$ for the corresponding G-equivariant line bundle on $\mathcal{B}$, and we set $\mathcal{F}(\lambda) = \mathcal{F} \otimes_{\mathcal{O}_{\mathcal{B}}} \mathcal{O}(\lambda)$ if $\mathcal{F} \in D(X)$ for some X mapping to $\mathcal{B}$.

Let W be the Weyl group, and set $W_{\mathrm{aff}} = W \ltimes \Lambda$. Then W, W_{aff} are Coxeter groups. Notice that W_{aff} is the affine Weyl group of the *Langlands dual* group ${}^L G$.

[1] The definition in *loc. cit.* is wider, we use a version convenient for our exposition.

It was mentioned above that $\widetilde{\mathfrak{g}}^{\mathrm{reg}} \cong \mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$; thus W acts on this space via its action on the second factor. The formulas $\Lambda \ni \lambda\colon \mathcal{F} \mapsto \mathcal{F}(\lambda)$, $W \ni w\colon \mathcal{F} \mapsto w^*(\mathcal{F})$ are easily shown to define an action[2] of W_{aff} on the category of coherent sheaves on $\widetilde{\mathfrak{g}}^{\mathrm{reg}}$.

The characterization of our "noncommutative Springer resolution" relies on the possibility to extend this action to a weaker structure on the whole of $\widetilde{\mathfrak{g}}$. To describe this weaker structure recall that to each Coxeter group one can associate an Artin braid group; let B_{aff} denote the group corresponding to W_{aff}. It admits a topological interpretation, as the fundamental group of the space of regular semi-simple conjugacy classes in the universal cover of the dual group ${}^LG(\mathbb{C})$. For $w \in W_{\mathrm{aff}}$ consider the minimal decomposition of w as a product of simple reflection, and take the product of corresponding generators of B_{aff}. This product is well known to be independent of the choice of the decomposition of w, thus we get a map $W_{\mathrm{aff}} \to B_{\mathrm{aff}}$ which is a one-sided inverse to the canonical surjection $B_{\mathrm{aff}} \to W_{\mathrm{aff}}$. We denote this map by $w \mapsto \widetilde{w}$. The map is not a homomorphism, however, we have $\widetilde{uv} = \widetilde{u} \cdot \widetilde{v}$ for any $u, v \in W_{\mathrm{aff}}$ such that $\ell(uv) = \ell(u) + \ell(v)$, where $\ell(w)$ denotes the length of the minimal decomposition of w. Let $B_{\mathrm{aff}}^+ \subset B_{\mathrm{aff}}$ be the sub-monoid generated by $\widetilde{w}$, $w \in W_{\mathrm{aff}}$.

For a simple reflection $s_\alpha \in W$ let $\boldsymbol{S}_\alpha \subset \widetilde{\mathfrak{g}}^2$ be the closure of the graph of s_α acting on $\widetilde{\mathfrak{g}}^{\mathrm{reg}}$. We let S_α denote the intersection of $\boldsymbol{S}_\alpha$ with $\widetilde{\mathcal{N}}^2$. Let $\mathbf{pr}_i^\alpha\colon \boldsymbol{S} \to \widetilde{\mathfrak{g}}$, $\mathrm{pr}_i^\alpha\colon S \to \widetilde{\mathcal{N}}$, where $i = 1, 2$, be the projections.

Let $\Lambda^+ \subset \Lambda$ be the set of dominant weights in Λ.

For a scheme Y over $\mathfrak{g}$ we set $\widetilde{Y} = \widetilde{\mathcal{N}} \times_{\mathfrak{g}} Y$, $\widetilde{\boldsymbol{Y}} = \widetilde{\mathfrak{g}} \times_{\mathfrak{g}} Y$.

Theorem 2.1. (a) *There exists an* (*obviously unique*) *action of* B'_{aff} *on* $D(\widetilde{\mathfrak{g}})$, $D(\widetilde{\mathcal{N}})$ *such that for* $\lambda \in \Lambda^+ \subset \Lambda \subset W'_{\mathrm{aff}}$ *we have* $\tilde{\lambda}\colon \mathcal{F} \mapsto \mathcal{F}(\lambda)$ *and for a simple reflection* $s_\alpha \in W$ *we have* $\tilde{s}_\alpha\colon \mathcal{F} \mapsto (\mathbf{pr}_1^\alpha)^*(\mathbf{pr}_2^\alpha)_*\mathcal{F}$ (*respectively,* $\tilde{s}_\alpha\colon \mathcal{F} \mapsto (\mathrm{pr}_1^\alpha)^*(\mathrm{pr}_2^\alpha)_*\mathcal{F}$).

(b) *This action induces an action on* $D(\widetilde{\boldsymbol{Y}})$, $D(\widetilde{Y})$ *for any scheme* Y *over* $\mathfrak{g}$ *such that* $\mathrm{Tor}_i^{\mathcal{O}_{\mathfrak{g}}}(\mathcal{O}_{\widetilde{\mathfrak{g}}}, \mathcal{O}_Y) = 0$, *respectively* $\mathrm{Tor}_i^{\mathcal{O}_{\mathfrak{g}}}(\mathcal{O}_{\widetilde{\mathcal{N}}}, \mathcal{O}_Y) = 0$, *for* $i > 0$.

Comment on the proof. The theorem can be deduced from material of either Section 3 or 4 below.

Remark 2.2. An example of Y satisfying the assumptions of the theorem is given by a transversal slice to a nilpotent orbit. In particular, if Y is a transversal slice to a subregular orbit, then $\widetilde{\mathcal{N}} \times_{\mathfrak{g}} Y$ is well known to be the minimal resolution of a simple surface singularity. The affine braid group action in this case coincides with the one constructed by Bridgeland in [27].

Remark 2.3. The induced action of B_{aff} on the Grothendieck group $K(\widetilde{\mathcal{N}})$ factors through W_{aff}. If one passes to the category of sheaves equivariant with respect to

[2]Throughout the paper by an action of a group on a category I mean *a weak action*, i.e., a homomorphism to the group of isomorphism classes of autoequivalences. I believe that in all the examples in this text a finer structure can be established, though I have not studied this question.

the multiplicative group, acting by dilations in the fibers of the projection $\widetilde{\mathcal{N}} \to \mathcal{B}$, then the induced action factors through the *affine Hecke algebra* $\mathcal{H}$, cf. discussion after Theorem 4.2. Furthermore, this construction yields an action of $\mathcal{H}$ on the Grothendieck group $K(\pi^{-1}(e))$ for each $e \in \mathcal{N}$; these $\mathcal{H}$ modules are called the standard $\mathcal{H}$-modules. Thus the theorem provides a *categorification* of the standard modules for the affine Hecke algebra.

The next result, which plays an important technical role in the proofs, is a categorical counterpart of the quadratic relation in the affine Hecke algebra, see discussion after Theorem 4.2.

Proposition 2.4. *For every simple reflection $s_\alpha \in W_{\mathrm{aff}}$ and every $\mathcal{F} \in \mathfrak{D}$ we have a (canonical) isomorphism in the quotient category $\mathfrak{D}/\langle \mathcal{F} \rangle$*

$$\tilde{s}_\alpha(\mathcal{F}) \cong \tilde{s}_\alpha^{-1}(\mathcal{F}) \mod \langle \mathcal{F} \rangle.$$

Here $\langle \mathcal{F} \rangle$ denotes the full triangulated subcategory generated by $\mathcal{F}$.

2.1.3. The t-structure and the noncommutative resolution. We will describe certain noncommutative resolutions A, $\boldsymbol{A}$ of $\mathcal{N}$, $\mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$ respectively, together with equivalences $D(A) \cong D(\widetilde{\mathcal{N}})$, $D(\boldsymbol{A}) \cong D(\widetilde{\mathfrak{g}})$, and show how they appear in representation theory. Such data is uniquely determined by the t-structures on $D(\widetilde{\mathcal{N}})$, $D(\widetilde{\mathfrak{g}})$, which are the images of the tautological t-structures on $D(A)$, $D(\boldsymbol{A})$.

Definition 2.5. Let D be a triangulated category equipped with an action of B_{aff}. A t-structure $(D^{<0},\ D^{\geq 0})$ on D will be called *braid right exact* if any $b \in B_{\mathrm{aff}}^+$ sends $D^{<0}$ to $D^{<0}$.

Theorem 2.6. (a) *Let X be either $\widetilde{Y}$ or $\widetilde{\boldsymbol{Y}}$, where $Y \to \mathfrak{g}$ is as in Theorem* 2.1.

The category $D(X)$ admits a unique t-structure which is

(i) *braid right exact, and*

(ii) *compatible with the standard t-structure on the derived category of vector spaces under the functor of derived global sections $R\Gamma$.*

(b) *There exists a vector bundle $\mathcal{E}_X$ on X, such that the functor $\mathcal{F} \mapsto R\operatorname{Hom}(\mathcal{E}, \mathcal{F})$ is an equivalence between $D(X)$ and $D(A_X)$, sending the t-structure described in* (a) *to the tautological t-structure on $D(A_X)$; here $A_X = \operatorname{End}(\mathcal{E}_X)^{\mathrm{op}}$, where the upper index denotes the opposite ring.*

Moreover, there exists a vector bundle $\mathcal{E} = \mathcal{E}_{\widetilde{\mathfrak{g}}}$ on $\widetilde{\mathfrak{g}}$, such that for any X we can take $\mathcal{E}_X$ to be the pull-back of $\mathcal{E}$ to X.

Remark 2.7. It is clear from the definitions that if X is smooth, then A_X is a noncommutative resolution of $Y \times_{\mathfrak{g}} \mathcal{N}$ or $Y \times_{\mathfrak{h}/W} \mathfrak{h}$. In particular, for $Y = \mathfrak{g}$ we get $A = A_{\widetilde{\mathcal{N}}}$, $\boldsymbol{A} = A_{\widetilde{\mathfrak{g}}}$, which are the promised noncommutative resolutions of $\mathcal{N}$, $\mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$.

We will call the t-structures described in Theorem 2.6 the *exotic* t-structures, the objects of their heart will be called exotic sheaves.

Example 2.8. Let $G = \mathrm{SL}(2)$, thus $\widetilde{\mathcal{N}}$ is the total space of the line bundle $\mathcal{O}(-2)$ on $\mathbb{P}^1$, and $\widetilde{\mathfrak{g}}$ is the total space of the vector bundle $\mathcal{O}_{\mathbb{P}^1}(-1) \oplus \mathcal{O}_{\mathbb{P}^1}(-1)$. In this case we can set $\mathcal{E} \cong \mathcal{O}_{\widetilde{\mathfrak{g}}} \oplus \mathcal{O}_{\widetilde{\mathfrak{g}}}(1)$.

This t-structure on $D(\widetilde{\mathfrak{g}})$ appeared in Bridgeland's proof of the derived equivalence conjecture for varieties of dimension three [28]. More precisely, for a flop of three-folds X, $X' \mapsto Y$ Bridgeland constructs some noncommutative resolution of Y which is derived equivalent to both X and X'. The simplest example of a three-fold flop is as follows: $X = X' = \widetilde{\mathfrak{g}}$, $Y = \mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$ and the two maps X, $X' \to Y$ are $\widetilde{\pi}'$ and $\widetilde{\pi}'' = \iota \circ \widetilde{\pi}'$, where ι is an involution of $\mathfrak{g} \times_{\mathfrak{h}/W} \mathfrak{h}$ given by $(x, h) \mapsto (x, -h)$. The t-structure on $D(\widetilde{\mathfrak{g}})$ given by Bridgeland's construction applied to this flop turns out to coincide with the t-structure provided by Theorem 2.6.

Example 2.9. Let Y be a transversal slice to the subregular orbit. Thus Y is isomorphic to the quotient $\mathbb{A}^2/\Gamma$ for some finite subgroup $\Gamma \subset \mathrm{SL}(2)$. The fiber product $X = \widetilde{\mathcal{N}} \times_{\mathfrak{g}} Y$ is the minimal resolution of Y. It is well known that there exists a natural equivalence $D(X) \cong D^{\Gamma}(\mathbb{A}^2)$. The exotic t-structure coincides with the one induced from the tautological t-structure on $D^{\Gamma}(\mathbb{A}^2)$. Thus A_X is Morita equivalent to the smash product algebra $\Gamma \# \mathcal{O}(\mathbb{A}^2)$. This t-structure appears also in [26].

2.1.4. Parabolic version. One can also consider the partial flag varieties $\mathcal{P} = G/P$, where $P \subset G$ is a parabolic subgroup; thus $\mathcal{P}$ parametrizes parabolic subalgebras $\mathfrak{p} \subset \mathfrak{g}$ of a given type. There exist parabolic versions of the Grothendieck–Springer spaces: $\widetilde{\mathfrak{g}}_{\mathcal{P}} = \{\mathfrak{p} \in \mathcal{P},\ x \in \mathfrak{p}\}$ and $\widetilde{\mathcal{N}}_{\mathcal{P}} = T^*(\mathcal{P})$. We have a proper map $\pi_{\mathcal{P}} \colon \widetilde{\mathfrak{g}} \to \widetilde{\mathfrak{g}}_{\mathcal{P}}$, $(gB, x) \mapsto (gP, x)$. Also, the projection $G/B \to G/P$ induces a closed embedding $\iota_{\mathcal{P}} \colon \mathcal{B} \times_{\mathcal{P}} \widetilde{\mathcal{N}}_{\mathcal{P}} \hookrightarrow \widetilde{\mathcal{N}}$; we let $\mathrm{pr}_{\mathcal{B}}^{\mathcal{P}}$ denote the projection $\mathcal{B} \times_{\mathcal{P}} \widetilde{\mathcal{N}}_{\mathcal{P}} \to \widetilde{\mathcal{N}}_{\mathcal{P}}$.

The following result easily follows from the results of [20].

Theorem 2.10. (a) *There exists a unique t-structure on $D(\widetilde{\mathfrak{g}}_{\mathcal{P}})$, whose heart contains the image of exotic sheaves under the functor $R\pi_{\mathcal{P}*} \colon D(\widetilde{\mathfrak{g}}) \to D(\widetilde{\mathfrak{g}}_{\mathcal{P}})$.*

(b) *There exists a unique t-structure on $D(\widetilde{\mathcal{N}}_{\mathcal{P}})$, such that for any object $\mathcal{F}$ in its heart the object $(\iota_{\mathcal{P}*}(\mathrm{pr}_{\mathcal{P}}^{\mathcal{B}})^*\mathcal{F})(\rho)$ is an exotic sheaf.*

One also has induced nice t-structures on $D(Y \times_{\mathfrak{g}} \widetilde{\mathfrak{g}}_{\mathcal{P}})$, $D(Y \times_{\mathfrak{g}} \widetilde{\mathcal{N}}_{\mathcal{P}})$ for Y satisfying a Tor vanishing condition; we omit the details to save space.

Example 2.11. Let $G = \mathrm{SL}(n+1)$ and $\mathcal{P} = \mathbb{P}^n$. The heart of the t-structure on $\widetilde{\mathcal{N}}_{\mathcal{P}} = T^*\mathbb{P}^n$ has a projective generator $\bigoplus_{i=0}^{n} \mathcal{O}_{T^*\mathbb{P}^n}(-i)$. The heart of the t-structure on $\widetilde{\mathfrak{g}}_{\mathbb{P}^n}$ has a projective generator $\bigoplus_{i=0}^{n} \mathcal{O}_{\widetilde{\mathfrak{g}}_{\mathcal{P}}}(i)$.

2.1.5. Reformulation in terms of t-structure assigned to alcoves. A connected component of the complement to the coroot hyperplanes H_α in the dual space to the *real* Cartan algebra $\mathfrak{h}^*_{\mathbb{R}}$ is called an alcove; in particular, the *fundamental alcove* A_0 is the locus of points where all positive coroots take value between zero and one. Let Alc be the set of alcoves. For $A_1,\ A_2 \in \mathrm{Alc}$ we will say that A_1 lies above A_2 if for any

positive coroot $\check{\alpha}$ and $n \in \mathbb{Z}$, such that the affine hyperplane $H_{\check{\alpha},n} = \{\lambda \mid \langle \check{\alpha}, \lambda \rangle = n\}$ separates A_1 and A_2, A_1 lies above $H_{\alpha,n}$, while A_2 lies below $H_{\alpha,n}$, i.e. for $\mu \in A_2$, $\lambda \in A_1$ we have $\langle \alpha, \mu \rangle < n < \langle \alpha, \lambda \rangle$.

Lemma 2.12. *There exists a unique map* $\mathrm{Alc} \times \mathrm{Alc} \to B_{\mathrm{aff}}$, $(A_1, A_2) \mapsto b_{A_1,A_2}$, *such that*

(i) $b_{A_2,A_3} b_{A_1,A_2} = b_{A_1 A_3}$ *for any* $A_1, A_2, A_3 \in \mathrm{Alc}$;

(ii) $b_{A_1,A_2} = \widetilde{w}$, *provided that* A_2 *lies above* A_1. *Here* $w \in W_{\mathrm{aff}}$ *is such that* $w(A_1) = A_2$.

The following result is equivalent to Theorem 2.6.

Theorem 2.13. *Let* $X = \widetilde{Y}$ *or* $\widetilde{\boldsymbol{Y}}$, *where* Y *is as in Theorem* 2.1. *There exists a unique collection of* t*-structures indexed by alcoves,* $(D_A^{\leq 0}(X),\ D_A^{>0}(X))$ *such that:*

(1) (Normalization) *The derived global sections functor* $R\Gamma$ *is* t*-exact with respect to the* t*-structure corresponding to* A_0.

(2) (Compatibility with the braid action) *The action of the element* b_{A_1,A_2} *sends the* t*-structure corresponding to* A_1 *to the* t*-structure corresponding to* A_2.

(3) (Monotonicity) *If* A_1 *lies above* A_2, *then* $D_{A_1}^{>0}(X) \supset D_{A_2}^{>0}(X)$.

Remark 2.14. The exotic t-structure described in Theorem 2.6 is the one attached to the fundamental alcove A_0 by the construction of Theorem 2.13.

Remark 2.15. The data described in Theorem 2.13 resemble the ones obtained by Bridgeland in the course of description of the manifold of stability conditions on some derived categories of coherent sheaves. To enhance this point we mention a positivity property of the t-structure $(D_A^{\leq 0}(X),\ D_A^{>0}(X))$; such properties play a role in the definition of stability conditions [24].

It is easy to show that each of the above t-structures induces a t-structure on the full subcategory $D^f(X) \subset D(X)$ consisting of complexes whose cohomology sheaves have proper support. Let $\mathcal{A}_A = D_A^{\leq 0}(X) \cap D_A^{\geq 0}(X)$ be the heart of the t-structure, and set $\mathcal{A}_A^f = \mathcal{A}_A \cap D^f(X)$. It is easy to show that $\mathcal{A}_A^f$ consists of objects *of finite length* in $\mathcal{A}_A$.

Assume that $\boldsymbol{k} = \mathbb{C}$ and X is smooth. Recall that for a smooth complex variety X we have the Chern character map $K(D^f(X)) \to H_*^{\mathrm{BM}}(X)$, where H_*^{BM} stands for the Borel–Moore homology of the corresponding complex variety endowed with the classical topology. We have a perfect pairing between cohomology and Borel–Moore homology.

We have a well-known identification $\mathfrak{h}^* = H^2(\mathcal{B})$.

Proposition 2.16. *For* $A \in \mathrm{Alc}$, $\mathcal{F} \in \mathcal{A}_A^f$, $\mathcal{F} \neq 0$ *and* $x \in A \subset \mathfrak{h}_{\mathbb{R}}^* \subset H^2(\mathcal{B})$ *we have*

$$\langle \mathrm{ch}(\mathcal{F}), \mathrm{pr}^*(\exp(x)) \rangle > 0,$$

where $\exp(x) = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^{\dim \mathcal{B}}}{(\dim \mathcal{B})!}$, *and* pr *stands for the projection* $X \to \mathcal{B}$.

Finally, we describe compatibility of our t-structures with duality.

Let $\mathcal{S}$ denote the Grothendieck–Serre duality functor.

Proposition 2.17. *$\mathcal{S}$ sends $\mathcal{A}_A^f$ to $\mathcal{A}_{-A}^f$, where $-A$ denotes the alcove opposite to A.*

2.2. Equivariant category and mutations of exceptional sets. The categories $D(\widetilde{\mathcal{N}})$, $D(\widetilde{\mathfrak{g}})$ have equivariant versions $D^G(\widetilde{\mathcal{N}})$, $D^G(\widetilde{\mathfrak{g}})$. It turns out that these equivariant categories carry t-structures which are, on the one hand, closely related to the above t-structures on non-equivariant categories, and, on the other hand, admit a direct description in terms of generating exceptional sets in a triangulated category.

Until the end of 2.3.2 *we assume that* $\mathrm{char}(\boldsymbol{k}) = 0$.

2.2.1. Exceptional sets and mutations. Recall that an ordered set of objects $\nabla = \{\nabla^i,\ i \in I\}$ in a triangulated category is called *exceptional* if we have $\mathrm{Hom}^\bullet(\nabla^i, \nabla^j) = 0$ for $i < j$; $\mathrm{Hom}^n(\nabla^i, \nabla^i) = 0$ for $n \neq 0$, and $\mathrm{End}(\nabla^i) = \boldsymbol{k}$. A set $\Delta = \{\Delta_i,\ i \in I\}$ of objects is called *dual* to ∇ if $\mathrm{Hom}^\bullet(\Delta_i, \nabla^i) = \boldsymbol{k}$, and $\mathrm{Hom}^\bullet(\Delta_i, \nabla^j) = 0$ for $i \neq j$; it is exceptional provided ∇ is, where the order on Δ is defined to be opposite to that on ∇. Let ∇, Δ be two dual exceptional sets which generate a triangulated category $\mathfrak{D}$; assume that $\{j \mid j \leq i\}$ is finite for every $i \in I$. Then there exists a unique t-structure $(\mathfrak{D}^{\geq 0}, \mathfrak{D}^{<0})$ on $\mathfrak{D}$, such that $\nabla \subset \mathfrak{D}^{\geq 0}$; $\Delta \subset \mathfrak{D}^{\leq 0}$. This construction is closely related to the definition of a perverse sheaf, see [14] for details.

Let $(I, \preceq)$ be an ordered set, and $\nabla^i \in \mathfrak{D}$, $i \in I$ be an exceptional set. Let $\leq$ be another order on I; we assume that $\{j \mid j \leq i\}$ is finite for every $i \in I$. We let $\mathfrak{D}_{\leq i}$ be the full triangulated subcategory generated by ∇^j, $j \leq i$, and similarly for $\mathfrak{D}_{<i}$. Then for $i \in I$ there exists a unique (up to a unique isomorphism) object ∇^i_{mut} such that $\nabla^i_{\mathrm{mut}} \in \mathfrak{D}_{\leq i} \cap \mathfrak{D}^{\perp}_{<i}$, and $\nabla^i_{\mathrm{mut}} \cong \nabla^i \mod \mathfrak{D}_{<i}$ (see e.g. [14]). The objects ∇^i_{mut} form an exceptional set indexed by $(I, \leq)$.

We will say that the exceptional set $(\nabla^i_{\mathrm{mut}})$ is the $\leq$ mutation of (∇^i). This construction is related, cf. [14], to the action of the braid group on the set of exceptional sets in a given triangulated category constructed in [21], this action is also called the action by mutations.

2.2.2. Exceptional sets in $D^G(\widetilde{\mathcal{N}})$. Recall the standard partial order $\preceq$ on the set Λ of weights of G, which is given by: $\lambda \preceq \mu$ if $\mu - \lambda$ is a sum of positive roots. Then line bundles $\mathcal{O}_{\widetilde{\mathcal{N}}}(\lambda)$ generate $D^G(\widetilde{\mathcal{N}})$, and we have $\mathrm{Hom}^\bullet(\mathcal{O}(\lambda), \mathcal{O}(\mu)) = 0$ unless $\mu \preceq \lambda$ and $\mathrm{Hom}^\bullet(\mathcal{O}(\lambda), \mathcal{O}(\lambda)) = \boldsymbol{k}$ [14]. Thus for any complete order on Λ compatible with the partial order $\preceq$, the set of objects $\mathcal{O}(\lambda)$ indexed by Λ with this order is an exceptional set generating $D^G(\widetilde{\mathcal{N}})$.

We now introduce another partial ordering $\leq$ on Λ. To this end, recall the 2-sided Bruhat partial order on the affine Weyl group W_{aff}. For $\lambda \in \Lambda$ let w_λ be the minimal length representative of the coset $W\lambda \subset W_{\mathrm{aff}}$. We set $\mu \leq \lambda$ if w_μ precedes w_λ in the Bruhat order.

We fix a complete order $\leq_{\mathrm{compl}}$ on Λ compatible with $\leq$; we assume that $\{\mu \mid \mu \leq_{\mathrm{compl}} \lambda\}$ is finite for any λ. We define the exceptional set ∇_λ to be the $\leq_{\mathrm{compl}}$ mutation of the set $\mathcal{O}(\lambda)$. It follows from the above that ∇^λ is an exceptional set generating $D^G(\widetilde{\mathcal{N}})$. We define the *equivariant exotic t-structure* to be the t-structure of the exceptional set ∇^λ, the objects in the heart will be called equivariant exotic sheaves.

We now state compatibility between exotic and equivariant exotic t-structures. Roughly speaking, over an orbit in $\mathcal{N}$ of codimension $2d$ they differ by a shift by d. To state this property more precisely, we need to recall the *perverse coherent t-structure* [13]. Let H be an algebraic group (assumed for simplicity of statements connected) acting on an algebraic variety X. Let $\boldsymbol{p}$ be a function, called the perversity function, from the set of H-invariant points of the scheme X to $\mathbb{Z}$. We assume that $\boldsymbol{p}$ is *strictly monotone and comonotone*, i.e. for points $x,\ y$, such that x lies in the closure of y we have $\boldsymbol{p}(y) < \boldsymbol{p}(x) < \boldsymbol{p}(y) + \dim(y) - \dim(x)$. Then one can define the perverse t-structure on $D^H(X)$, which shares some properties with perverse t-structure on the derived category of constructible sheaves [7]. For example, each perverse coherent sheaf (i.e., object in the heart of the t-structure) has finite length, and irreducible objects are in bijection with pairs $(O, \mathcal{L})$, where $O \subset X$ is an H-orbit, and $\mathcal{L}$ is an irreducible H-equivariant vector bundle on O. In particular, if the action is such that all orbits have even dimension, then the perversity function $\boldsymbol{p}(x) = \frac{\operatorname{co}\dim x}{2}$, called the middle perversity, is strictly monotone and comonotone. It is well-known that the adjoint action of a semi-simple group G on the nil-cone $\mathcal{N}$ has even dimensional orbits.

This construction works also for the category $D^H(A)$, where A is a coherent sheaf of associative $\mathcal{O}_X$ algebras equivariant under H.

Proposition 2.18. *There exists a G-equivariant vector bundle $\mathcal{E}$ on $\widetilde{\mathcal{N}}$, such that $\mathcal{E}$, with the G-equivariant structure forgotten, is a projective generator for the heart of the exotic t-structure.*

We have an equivalence $\mathcal{F} \mapsto R\operatorname{Hom}(\mathcal{E}, \mathcal{F})$ between $D^G(\widetilde{\mathcal{N}})$ and $D^G(A)$, where $A = \operatorname{End}(\mathcal{E})^{\mathrm{op}}$. Under this equivalence the equivariant exotic t-structure corresponds to the perverse coherent t-structure of the middle perversity.

2.3. Grading on exotic sheaves and canonical bases

2.3.1. Graded equivariant category and positivity by Frobenius weights. We proceed to state a deep property of exotic sheaves related to an additional grading on the Ext spaces between them. Recall the current assumption that $\operatorname{char}(\boldsymbol{k}) = 0$.

Consider the category $D^{G\times\mathbb{G}_m}(\widetilde{\mathcal{N}})$, where $\mathbb{G}_m$ acts on $\widetilde{\mathcal{N}}$ by $t\colon x \mapsto t^2 x$. For $d \in \mathbb{Z}$ let $\mathcal{F} \mapsto \mathcal{F}(d)$ denote twisting by the d-th power of the tautological character of $\mathbb{G}_m$. We refer to [14] for an elementary description of a canonical lifting $\widetilde{\Delta}_\lambda$, $\widetilde{\nabla}^\lambda$ of Δ_λ, ∇^λ to $D^{G\times\mathbb{G}_m}$. This also fixes a lifting $\tilde{L}$ of each irreducible equivariant exotic sheaf L to $D^{G\times\mathbb{G}_m}$.

Theorem 2.19. *For irreducible exotic equivariant sheaves L_1, L_2 we have*

$$\mathrm{Ext}^i(\tilde{L}_1, \tilde{L}_2(d)) = 0$$

for $d \leq 0$ and all i.

Remark 2.20. The theorem follows from results of [12] on relation between exotic sheaves and perverse sheaves on the affine flag manifold of the dual group, see also Proposition 4.5 below. They allow to deduce the theorem from Gabber's Theorem [7] on positivity of weights of Frobenius action on Ext's between pure perverse sheaves of the same weight. Thus it is the least elementary of the results mentioned so far in this text.

The motivation for the theorem is its consequence below, which shows (in most cases) that classes of exotic sheaves form a *canonical basis* in the Grothendieck group. This is parallel to the proof of the Kazhdan–Lusztig conjecture: according to Soergel, cf. [50], the latter is equivalent to the statement that for a certain explicitly defined graded version of Bernstein–Gel'fand–Gel'fand category O the grading on Ext^1 between irreducible objects has vanishing components of non-positive degrees. The only known way to prove this vanishing is to identify category O with a category of perverse sheaves or Hodge D-modules, and use deep information about purity of Frobenius or Hodge weights.

Remark 2.21. Another application of Theorem 2.19 is explained in [14]. Together with the *Koszul duality* formalism of [9] it allows one to show that equivariant exotic sheaves control cohomology of quantum groups at a root of unity with coefficients in a tilting module.

2.3.2. Non-equivariant graded category and canonical bases. We fix $X = \widetilde{\mathcal{N}}$. Recall the category $\mathcal{A}^f = \mathcal{A}^f_{A_0} \subset \mathcal{A}$ of exotic sheaves of finite length. It is easy to see that $\mathcal{A}^f = \bigoplus_{e \in \mathcal{N}} \mathcal{A}_e$, where $\mathcal{A}_e = \mathcal{A} \cap \mathfrak{D}_e$, and $\mathfrak{D}_e \subset D(\widetilde{\mathcal{N}})$ is the full subcategory of complexes whose cohomology sheaves are set-theoretically supported on $\mathcal{B}_e = \pi^{-1}(e)$. We have $K(\mathcal{A}_e) \cong K(\mathcal{B}_e)$. Furthermore, the Chern character map provides an isomorphism $K(\mathcal{B}_e)_F \cong H^{\mathrm{BM}}_*(\mathcal{B}_e)_F$, where F denotes a coefficient field of characteristic zero ($\mathbb{C}$ or $\overline{\mathbb{Q}_l}$), see, e.g, [19].

The classes of irreducible objects form a basis in $K(\mathcal{A}_e)$. We proceed to explain the properties of the category, which are needed to relate this basis to *the canonical bases* in $H^{\mathrm{BM}}_*(\mathcal{B}_e)$. The definition of the latter is due to Lusztig [45], and follows the example of Kashiwara's characterization of crystal bases [42]. More precisely, Lusztig suggested a way to characterize a basis in $H^{\mathrm{BM}}_*(\mathcal{B}_e)$, and conjectured that a basis satisfying his axioms exists; he showed that it is then unique (up to a sign). We will not recall Lusztig's characterization in detail; instead we describe its structure and explain the properties of exotic sheaves, which imply (modulo a technicality, which is easy to check in many cases) that Lusztig's axioms are satisfied by the basis of irreducible exotic sheaves.

One can find a homomorphism $\varphi\colon \mathrm{SL}(2) \to G$, such that $d\varphi$ sends the standard upper triangular generator of $sl(2)$ to e. Then we get an action a_φ of the multiplicative group $\mathbb{G}_m$ on $\mathfrak{g}$ given $a_\varphi(t)\colon x \mapsto t^2 \cdot \mathrm{ad}(\varphi(\mathrm{diag}(t^{-1}, t)))x$. This action fixes e.

We let $\mathfrak{D}_e^{\mathbb{G}_m} \subset D^{\mathbb{G}_m}(\widetilde{\mathcal{N}})$ be the full subcategory of complexes, which are set theoretically supported on $\pi^{-1}(e)$. Twisting by the tautological character of $\mathbb{G}_m$ defines an auto-equivalence of this category, which we denote by $\mathcal{F} \mapsto \mathcal{F}(1)$. The exotic t-structure is inherited by the $\mathbb{G}_m$-equivariant category; we let $\mathcal{A}_e^{\mathrm{gr}}$ denote the heart of the latter. It is easy to see that the forgetful functor $\mathfrak{D}_e^{\mathbb{G}_m} \to \mathfrak{D}_e$ sends $\mathrm{Irr}(\mathcal{A}_e^{\mathrm{gr}})$ to $\mathrm{Irr}(\mathcal{A}_e)$, where Irr stands for the set of isomorphism classes of irreducible objects. This gives a bijection $\mathrm{Irr}(\mathcal{A}_e^{\mathrm{gr}})/\mathbb{Z} \cong \mathrm{Irr}(\mathcal{A}_e)$, where $\mathbb{Z}$ acts on $\mathrm{Irr}(\mathcal{A}_e^{\mathrm{gr}})$ by $\mathcal{F} \mapsto \mathcal{F}(n)$. We also have $K(\mathcal{A}_e^{\mathrm{gr}}) \cong K^{\mathbb{G}_m}(\mathcal{B}_e) \cong K(\mathcal{B}_e)[v, v^{-1}]$, where multiplication by v corresponds to twisting by the tautological character of $\mathbb{G}_m$.

The canonical basis in $K(\mathcal{B}_e)[v, v^{-1}]$ is characterized (up to a sign) by two properties: *invariance under an involution* and *asymptotic orthogonality* [45]. These are reflected, respectively, in categorical properties (i) and (ii) in the next theorem.

Notice that the action of B_{aff} on $\mathfrak{D}_e$ is inherited by $\mathfrak{D}_e^{\mathbb{G}_m}$. Recall that $\mathcal{S}$ is the Grothendieck–Serre duality.

In view of Theorem 2.13 and Proposition 2.17, the contravariant auto-equivalence $\widetilde{w}_0 \circ \mathcal{S}$ is t-exact with respect to the t-structure corresponding to the fundamental alcove A_0, hence it permutes irreducible objects of $\mathcal{A}_{A_0}$; here $w_0 \in W$ is the long element.

Theorem 2.22. *There exists a canonical section of the map* $\mathrm{Irr}(\mathcal{A}_e^{\mathrm{gr}}) \to \mathrm{Irr}(\mathcal{A}_e)$, $L \mapsto \tilde{L}$, *such that*

(i) *The image of the section is invariant under every automorphism of G which is identity on the image of φ_e, and also under $\widetilde{w}_0 \circ \mathcal{S}$.*

(ii) $\mathrm{Ext}^1_{\mathcal{A}_e^{\mathrm{gr}}}(\tilde{L}_1, \tilde{L}_2(i)) = 0$ *for* $i \le 0$ *and any* $L_1,\ L_2 \in \mathrm{Irr}(\mathcal{A}_e)$*; here* $\bar{\mathcal{A}}_e^{\mathrm{gr}} \subset \mathcal{A}_e$ *is the full subcategory of objects where the ideal of the point e in $\mathcal{O}(\mathfrak{g})$ acts by zero.*

Comments on the proof. The theorem can be deduced formally from Theorem 2.19 and Proposition 2.18. Thus its proof relies on ideas of geometric Langlands duality used in [2], and on Gabber's Theorem (see comments after Theorem 2.19).

Corollary 2.23. *Suppose that the involution $\tilde{\beta}$ defined in* [45], §5.11 *induces identity on the specialization at $q = 1$. Then Conjecture* 5.12 *of loc. cit., except, possibly,* 5.12 (g)*, holds; moreover, the signed basis $\boldsymbol{B}^{\pm}_{\mathcal{B}_e}$, whose existence is conjectured in loc. cit., is formed by the classes of the objects $\tilde{L}$, where L runs over irreducible objects in $\mathcal{A}_e$.*

Remark 2.24. The assumptions of the corollary are easy to check in many cases, e.g., if the nilpotent element e is regular in a Levi subalgebra.

Remark 2.25. In fact, in [45] Lusztig works with sheaves which are also equivariant under a maximal torus in the centralizer of e. We omit this version here to simplify notations, treating this set-up does not involve new ideas.

Remark 2.26. Validity of Conjecture 5.12 (g) of [45] is related to the following question. Let $Y \subset \mathfrak{g}$ be a (Slodowy) transversal slice to a G orbit in $\mathcal{N}$, and $X = \widetilde{\mathcal{N}} \times_{\mathfrak{g}} Y$. Let A_X be as in Theorem 2.6. One can show that A can be endowed with a natural grading; moreover, Theorem 2.22 is equivalent to the fact that this grading can be chosen so that the graded components of negative degree vanish, while the component of degree zero is semi-simple. The question is whether the resulting graded algebra is Koszul. If e is subregular, then the positive answer is easy to prove.

2.3.3. Independence of the (large) prime. It is not hard to show that (co)homology of the Springer fiber is independent of the ground field $\boldsymbol{k}$, i.e. we have canonical isomorphisms $H^{\mathrm{BM}}_\bullet(\mathcal{B}^{\boldsymbol{k}}_e) \cong H^{\mathrm{BM}}_\bullet(\mathcal{B}^{\mathbb{C}}_e)$, where the upper index denotes the ground field, and $H^{\mathrm{BM}}_\bullet$ stands for l-adic Borel–Moore homology, $l \neq \mathrm{char}(\boldsymbol{k})$.

The definition of the exotic t-structure is not specific to a particular ground field. This allows one to prove the following.

Proposition 2.27. *For all but finitely many prime numbers p the following is true. The classes in $H^{\mathrm{BM}}_\bullet(\mathcal{B}^{\boldsymbol{k}}_e) = H^{\mathrm{BM}}_\bullet(\mathcal{B}^{\mathbb{C}}_e)$ of irreducible exotic sheaves over $\boldsymbol{k}$ of characteristic p coincide with the classes of irreducible exotic sheaves over $\mathbb{C}$.*

3. D-modules in positive characteristic and localization theorem

3.1. Generalities on crystalline D-modules in positive characteristic

3.1.1. Definition and description of the center. Let X be a smooth variety over the field $\boldsymbol{k}$.

The sheaf $\mathcal{D} = \mathcal{D}_X$ of *crystalline differential operators* (or differential operators without divided powers, or PD differential operators) on X is defined as the enveloping of the tangent Lie algebroid, i.e., for an affine open $U \subset X$ the algebra $\mathcal{D}(U)$ contains the subalgebra $\mathcal{O}$ of functions, has an $\mathcal{O}$-submodule identified with the Lie algebra of vector fields $\mathrm{Vect}(U)$ on U, and these subspaces generate $D(U)$ subject to relations $\xi_1\xi_2 - \xi_2\xi_1 = [\xi_1, \xi_2] \in \mathrm{Vect}(U)$ for $\xi_1, \xi_2 \in \mathrm{Vect}(U)$, and $\xi \cdot f - f \cdot \xi = \xi(f)$ for $\xi \in \mathrm{Vect}(U)$ and $f \in \mathcal{O}(U)$.

If $\mathrm{char}(\boldsymbol{k}) = 0$, then $\mathcal{D}_X$ is the familiar sheaf of differential operators. From now on assume that $\boldsymbol{k}$ is of characteristic $p > 0$. Then $\mathcal{D}_X$ shares some features with the characteristic zero case; for example, $\mathcal{D}_X$ carries an increasing filtration "by order of a differential operator", and the associated graded $\mathrm{gr}(\mathcal{D}_X) \cong \mathcal{O}_{T^*X}$ canonically. On the other hand, some phenomena are special to the characteristic p setting. We have an action map $\mathcal{D}_X \to \mathrm{End}(\mathcal{O}_X)$, which is not injective, unlike in the case of characteristic zero. For example, if $X = \mathbb{A}^1 = \mathrm{Spec}(k[x])$, the section $\partial_x^p \neq 0$ of $\mathcal{D}_X$ acts by zero on $\mathcal{O}$. Also, $\mathcal{D}_X$ has a huge center; for example, if $X = \mathbb{A}^n = \mathrm{Spec}(k[x_1, \ldots, x_n])$, then x_i^p and $\partial_{x_i}^p$ are readily seen to generate the center $Z(\mathcal{D}_{\mathbb{A}^n})$ freely. More generally, for any X the center $Z(\mathcal{D}_X)$ is freely generated by elements of the form f^p, $f \in \mathcal{O}_X$

and $\xi^p - \xi^{[p]}$, $\xi \in \mathrm{Vect}_X$, where $\xi^{[p]}$ is the *restricted power* of the vector field ξ; it is characterized by $\mathrm{Lie}_{\xi^{[p]}}(f) = \mathrm{Lie}^p_\xi(f)$ for $f \in \mathcal{O}_X$, where Lie stands for the Lie derivative. The center $Z(\mathcal{D}_X)$ is canonically isomorphic to the sheaf of rings $\mathcal{O}_{T^*X^{(1)}}$ where the super-index (1) stands for Frobenius twist.[3] Thus $\mathcal{D}_X$ can be considered as a quasi-coherent sheaf of algebras on $T^*X^{(1)}$.

3.1.2. Azumaya property. Recall that an *Azumaya algebra* on a scheme X is a locally free sheaf $\mathcal{A}$ of associative $\mathcal{O}_X$ algebras, such that the fiber of $\mathcal{A}$ at every geometric point is isomorphic to a matrix algebra. The following fundamental observation is due to Mirković and Rumynin, though a weak form of it can be traced to an earlier work [40].

Theorem 3.1 ([19]). *$\mathcal{D}_X$ is an Azumaya algebra of rank $p^{2\dim(X)}$ on $T^*X^{(1)}$.*

See [48], [10] for generalizations and applications.

Recall that two Azumaya algebras $\mathcal{A}$, $\mathcal{A}'$ are called equivalent (we then write $\mathcal{A} \sim \mathcal{A}'$) if they are Morita equivalent, i.e. if there exists a coherent locally projective sheaf $\mathcal{M}$ of $\mathcal{A} - \mathcal{A}'$ bimodules, such that $\mathcal{A}' \xrightarrow{\sim} \mathrm{End}(\mathcal{M})^{\mathrm{op}}$; we will then say that $\mathcal{M}$ provides an equivalence between $\mathcal{A}$ and $\mathcal{A}'$. In particular, an Azumaya algebra $\mathcal{A}$ is *split* if $\mathcal{A} \sim \mathcal{O}_X$; this happens iff $\mathcal{A} \cong \mathrm{End}(\mathcal{E})$ for a vector bundle $\mathcal{E}$. For two equivalent Azumaya algebras $\mathcal{A}$, $\mathcal{A}'$ we have an equivalence of categories of modules $\mathrm{Coh}(X, \mathcal{A}) \cong \mathrm{Coh}(X, \mathcal{A}')$, depending on the choice of a bimodule providing the equivalence between $\mathcal{A}$ and $\mathcal{A}'$; in particular, for a split Azumaya algebra we have $\mathrm{Coh}(X, \mathcal{A}) \cong \mathrm{Coh}(X)$.

For a smooth variety X over a positive characteristic field, the Azumaya algebra $\mathcal{D}_X$ is not split unless $\dim(X) = 0$. However, it is split on the zero section, see [48] for more information.

We will also need a twisted version of differential operators. If $\mathcal{L}$ is a line bundle on X, then one can consider the sheaf $\mathcal{D}^{\mathcal{L}} = \mathcal{D}^{\mathcal{L}}_X$ of differential operators in $\mathcal{L}$. A similar argument shows that this is also an Azumaya algebra over $T^*X^{(1)}$; moreover, we have a canonical equivalence

$$\mathcal{D}_X \sim D^{\mathcal{L}}_X, \tag{1}$$

given by the bimodule $D_X \otimes_{\mathcal{O}(X)} \mathcal{L}^{-1}$.

Remark 3.2. Notice that if $\mathcal{L} = \mathcal{L}_0^p = \mathrm{Fr}^*(\mathcal{L}_0)$ for some line bundle $\mathcal{L}_0$, then we have a canonical isomorphism $\mathcal{D}^{\mathcal{L}}_X \cong \mathcal{D}_X$; however, the above equivalence $\mathcal{D}^{\mathcal{L}}_X \cong \mathcal{D}_X$ is not identity, but rather tensor product *over the ring* $\mathcal{O}^p_X = \mathcal{O}_{X^{(1)}}$ with the line bundle ${\mathcal{L}_0}^{(1)}$.

[3] Recall that Frobenius twist of a variety X over a perfect field $\boldsymbol{k}$ is defined to be isomorphic to X as an abstract scheme, with the $\boldsymbol{k}$-linear structure twisted by Frobenius. Not only $X \cong X^{(1)}$ as abstract schemes, but also $X \cong X^{(1)}$ as $\boldsymbol{k}$-schemes, provided that X is defined over $\mathbb{F}_p$. For this reason we will sometimes identify X with $X^{(1)}$ and omit Frobenius twist from notation.

3.2. Crystalline operators on $\mathcal{B}$. We now consider $X = \mathcal{B}$. We abbreviate $\mathcal{D}_{\mathcal{B}}^{\mathcal{O}(\lambda)} = \mathcal{D}^{\lambda}$.

3.2.1. Splitting the Azumaya algebra. It was mentioned above that $\mathcal{D}_{\mathcal{B}}$ splits on the zero section. In fact, we have the following stronger statement.

Theorem 3.3. (a) *There exists an Azumaya algebra $\mathcal{A}$ on $\mathcal{N}^{(1)}$, such that $\mathcal{D}^{-\rho} \cong \pi^*(\mathcal{A})$.*

(b) *For any λ we have an equivalence of Azumaya algebras on $\mathcal{N}^{(1)}$: $\mathcal{D}^{\lambda} \sim \pi^*(\mathcal{A})$.*

(c) *$\mathcal{D}^{\lambda}(\mathcal{B})$ is split on the formal neighborhood of every fiber of π.*

Sketch of proof. (a) reduces to irreducibility of baby Verma modules with highest weight $-\rho$, which follows from [29]. It implies, moreover, that the statement holds for $\mathcal{A}$ being the quotient of the enveloping algebra $U(\mathfrak{g})$ by the central ideal corresponding to $-\rho$. (b) follows from (a) in view of the equivalence (1). Finally, (c) follows from (b), since every Azumaya algebra over a complete local ring with an algebraically closed residue field splits. □

Let $\mathcal{D}^{\lambda}$-mod$^f \subset \mathrm{Coh}(\widetilde{\mathcal{N}}^{(1)}, \mathcal{D}^{\lambda})$ the full subcategory of sheaves, whose support (which is a subvariety in $\widetilde{\mathcal{N}}^{(1)}$) is proper. Let $\mathrm{Coh}^f(\widetilde{\mathcal{N}}) \subset \mathrm{Coh}(\widetilde{\mathcal{N}})$ be the full subcategory of sheaves with proper support.

Corollary 3.4. *For every $\lambda \in \Lambda$ we have an equivalence $\mathcal{D}^{\lambda}$-mod$^f \cong \mathrm{Coh}^f(\widetilde{\mathcal{N}})$.*

Proof. Since the target of π is affine, a subscheme Z in $\widetilde{\mathcal{N}}^{(1)}$ is proper iff it lies in a finite union of nilpotent neighborhoods of Springer fibers. Thus the claim follows from Theorem 3.3(b). □

For each $\lambda \in \Lambda$ and $e \in \mathcal{N}$ we fix the splitting bundle $\mathcal{E}_e^{\lambda}$ for $\mathcal{D}^{\lambda}$ on the formal neighborhood of $\pi^{-1}(e)$ as follows. For $\lambda = -\rho$ we let $\mathcal{E}_e^{\lambda}$ be the pull-back under π of a splitting bundle for the Azumaya algebra $\mathcal{A}$ on the formal neighborhood of e in $\mathcal{N}^{(1)}$. For a general λ we get $\mathcal{E}^{\lambda}$ from $\mathcal{E}^{-\rho}$ by applying the canonical equivalence (1) between $\mathcal{D}^{-\rho}$ and $\mathcal{D}^{\lambda}$; thus $\mathcal{E}^{\lambda} = \mathcal{E}^{-\rho} \otimes_{\mathcal{O}_{\mathcal{B}}} \mathcal{O}(\lambda + \rho)$.

We let F_{λ} denote the resulting equivalence between $\mathcal{D}^{\lambda}$-mod^f and $\mathrm{Coh}^f(\widetilde{\mathcal{N}})$. Notice that for $\lambda' = \lambda + p\mu$ the sheaves of algebras $\mathcal{D}^{\lambda}$ and $\mathcal{D}^{\lambda'}$ are canonically identified; however, the equivalences F_{λ} and $F_{\lambda'}$ are different, cf. Remark 3.2.

3.2.2. Derived localization in positive characteristic. Let $U = U(\mathfrak{g})$ be the enveloping algebra.

Assume first that $\mathrm{char}(\boldsymbol{k}) = 0$. Recall the famous Localization Theorem [5], [30], which provides an equivalence U^{λ}-mod $\cong \mathcal{D}^{\lambda}$-mod$(\mathcal{B})$, where λ is a dominant integral weight, $\mathcal{D}^{\lambda}$-mod denotes the corresponding twisted D-modules category, and U^{λ}-mod is the category of $\mathfrak{g}$-modules with central character corresponding to λ. For two integral weights λ, μ the categories $\mathcal{D}^{\mu}$-mod and $\mathcal{D}^{\lambda}$-mod can be identified by means of the equivalence $T_{\mu}^{\lambda}\colon \mathcal{F} \mapsto \mathcal{F} \otimes \mathcal{O}(\lambda - \mu)$. If λ, μ are dominant, then

the global sections functors intertwine this equivalence with the translation functor, which provides an equivalence U^λ-mod $\cong U^\mu$-mod.

Assume now that μ is integral regular, thus $\mu = w(\lambda+\rho)-\rho$ for some dominant integral λ, $w \in W$. Then the functor of global sections on D_μ-mod is no longer exact; however, it follows from [6] that the derived functor $R\Gamma = R\Gamma_\mu : D^b(\mathcal{D}^\mu\text{-mod}) \to D^b(U^\lambda\text{-mod})$ is still an equivalence. The triangle formed by the three equivalences $R\Gamma_\mu$, T^λ_μ, $R\Gamma_\lambda$ does not commute. Thus we get an auto-equivalence R_w of $D^b(\mathcal{D}^\lambda\text{-mod})$, $R_w = R\Gamma_\lambda^{-1} \circ R\Gamma_\mu \circ T^\mu_\lambda$. In [6] it is shown that R_w can be described by an explicit correspondence, which makes it natural to call R_w the *Radon transform*, or *the intertwining functor*. Moreover, the assignment $\widetilde{w} \mapsto R_w$ extends to an action of the Artin braid group B attached to G on $D^b(U^\lambda\text{-mod})$.

A part of this picture can be generalized to characteristic p.

The obvious characteristic p analogue of the above equivalence of abelian categories does not hold for any integral λ. Indeed, it is well known that for any coherent sheaf $\mathcal{F}$ on the (Frobenius twist of) a smooth variety the sheaf $\mathrm{Fr}^*(\mathcal{F})$ carries a flat connection; in particular, so does the sheaf $\mathrm{Fr}^*(L) = L^{\otimes p}$, where L is a line bundle. Thus for $\mathcal{F} \in \mathcal{D}^\lambda$-mod we have $\mathcal{F} \otimes L^p \in \mathcal{D}^\lambda$-mod. If L is anti-ample and the support of $\mathcal{F}$ is projective of positive dimension, then some of the higher derived functors $R^i\Gamma(\mathcal{F} \otimes L^{dp}) \neq 0$ for large d.

However, we do have an analogue of the "derived" localization theorem. From now on assume that $\mathrm{char}(\boldsymbol{k}) = p > 0$.

The center Z of U contains the subalgebra $Z_{\mathrm{HC}} = U^G \cong \mathrm{Sym}(\mathfrak{h})^W$, which we call the Harish-Chandra center. We have a natural map $\Lambda/p\Lambda \to \mathfrak{h}^*/W$, $\lambda \mapsto d\lambda$ mod W. Thus every $\lambda \in \Lambda$ defines a maximal ideal of Z_{HC}. We let $U^\lambda = U \otimes_{Z_{\mathrm{HC}}} \boldsymbol{k}$ denote the corresponding central reduction. Notice that the set of weights $\mu \in \Lambda$, such that the quotients U^λ and U^μ of U coincide, is precisely the W_{aff}-orbit of λ with respect to the action $w \bullet \lambda = p\, w\big(\frac{\lambda+\rho}{p}\big) - \rho$. We will say that $\lambda \in \Lambda$ is p-regular if the stabilizer in W of $\lambda + p\Lambda \in \Lambda/p\Lambda$ is trivial.

We also have another central subalgebra $Z_{\mathrm{Fr}} \subset U$, called the Frobenius center. It is generated by expressions of the form $x^p - x^{[p]}$, $x \in \mathfrak{g}$, where the *restricted power* map $x \mapsto x^{[p]}$ is characterized by $\mathrm{ad}(x^{[p]}) = \mathrm{ad}(x)^p$. Thus maximal ideals of Z_{Fr} are in bijection with points of $\mathfrak{g}^* \cong \mathfrak{g}$.

Let U^λ-mod denote the category of finitely generated U^λ-modules, and let $U^\lambda\text{-mod}^f \subset U^\lambda$-mod be the full subcategory of finite length modules.

For a pair $\lambda \in \Lambda$, $e \in \mathfrak{g}^*$ let $U^\lambda_{\hat{e}}$-mod be the category of finitely generated U^λ-modules, which are killed by some power of the maximal ideal of e in Z_{Fr}. This category is zero unless $e \in \mathcal{N}$. We also have $U^\lambda\text{-mod}^f = \bigoplus_{e\in\mathcal{N}} U^\lambda_{\hat{e}}\text{-mod}$.

Let $\mathcal{D}^\lambda_{\hat{e}}$-mod $\subset \mathcal{D}^\lambda$-mod be the full subcategory of objects which are supported on a nilpotent neighborhood of $\pi^{-1}(e)$; here we think of $\mathcal{D}^\lambda$ modules as sheaves on $\widetilde{\mathcal{N}}^{(1)}$ with an additional structure.

Theorem 3.5 ([19]). (a) *We have a natural isomorphism* $\Gamma(\mathcal{D}^\lambda) \cong U^\lambda$ *for every* $\lambda \in \Lambda$.

(b) *If $\lambda \in \Lambda$ is p-regular, then the derived global sections functor provides an equivalence $R\Gamma_\lambda \colon D(\mathcal{D}^\lambda) \xrightarrow{\sim} D(U^\lambda)$. It restricts to equivalences $D^b(\mathcal{D}^\lambda\text{-mod}^f) \cong D^b(U^\lambda\text{-mod}^f)$, $D^b(\mathcal{D}^\lambda_{\hat{e}}\text{-mod}) \cong D^b(U^\lambda_{\hat{e}}\text{-mod})$.*

Remark 3.6. This theorem has several versions and generalization. One can work with the more general categories of twisted D-modules, thereby obtaining a category of modules over an Azumaya algebra on the formal neighborhood of $\widetilde{\mathcal{N}}$ in $\widetilde{\mathfrak{g}}$, or more general subschemes or formal completions of $\widetilde{\mathfrak{g}}$. For singular weights λ there is a version of the theorem that relates derived categories of modules to sheaves on (the neighborhoods of) *parabolic Springer fibers* [20]. Another construction works with differential operators on a partial flag variety G/P for a parabolic subgroup $P \subset G$, *loc. cit.,* cf. also subSection 2.1.4 above.

For a scheme Y mapping to $\mathfrak{g}$ and satisfying the Tor vanishing conditions of Theorem 2.1 we have an equivalence between the derived category of modules over Azumaya algebras on $\widetilde{\boldsymbol{Y}}, \widetilde{Y}$ obtained as pull-back of the algebra of (twisted) differential operators and derived category of modules over the algebra of global sections.

If Y is a transversal slice to a nilpotent orbit, then the algebra of global sections of the Azumaya algebra on $\widetilde{Y}$ is probably related to Premet's quantization of Slodowy slices and generalized Whittaker D-modules, see [49], [38].

There exists a generalization of this result for Y not satisfying the Tor vanishing condition. It involves coherent sheaves on the *differential graded* scheme, which is the derived fiber product of Y and $\widetilde{\mathfrak{g}}$ over $\mathfrak{g}$. The particular case $Y = \{0\}$ is closely related to the description of the derived category of the principal block in representations of a quantum group at a root of unity provided by [3].

Proposition 3.7. (a) *A weight $\lambda \in \Lambda$ is p-regular iff $\frac{\lambda+\rho}{p}$ lies in some alcove.*

(b) *The t-structure on $\mathcal{D}^0$-mod induced by the equivalence $R\Gamma^\lambda \circ T_0^\lambda$ for a p-regular λ depends only on the alcove of $\frac{\lambda+\rho}{p}$ (see the beginning of Section 3.2.2 for notation).*

Thus we get a collection of t-structure on $\mathcal{D}^0$-mod indexed by alcoves; we denote the t-structure attached to $A \in \mathrm{Alc}$ by $D_A^{<0}(\mathcal{D})$, $D_A^{\geq 0}(\mathcal{D})$. The following properties of the collection follow from [19], [20].

Theorem 3.8. (a) *Let A_1, A_2 be two alcoves. If A_1 lies above A_2, then $D_{A_1}^{>0}(\mathcal{D}) \supset D_{A_2}^{>0}(\mathcal{D})$.*

(b) *There exists an action of B_{aff} on $D(\mathcal{D})$, such that the following holds. Let $\lambda \in \Lambda$, $\frac{\lambda+\rho}{p} = w\left(\frac{\rho}{p}\right)$ for $w \in W_{\mathrm{aff}}$, and let A be the alcove of $\frac{\lambda+\rho}{p}$. Then $R\Gamma_\lambda \cong R\Gamma_0 \circ b_{A_0,A_1}$. Thus b_{A_0,A_1} sends the t-structure $D_{A_0}^{<0}(\mathcal{D})$, $D_{A_0}^{\geq 0}(\mathcal{D})$ to the t-structure $D_A^{<0}(\mathcal{D})$, $D_A^{\geq 0}(\mathcal{D})$.*

(c) *The restriction of the B_{aff} action to $D^b(\mathcal{D}^\lambda - mod^f) \cong D^b(\mathrm{Coh}^f(\widetilde{\mathcal{N}}))$ coincides with the restriction of the action from Theorem 2.1.*

(d) *The derived global sections functor* $R\Gamma$ *on* $D^b(\mathrm{Coh}^f(\widetilde{\mathcal{N}}))$ *is* t*-exact with respect to the* t*-structure induced from* $(D^{<0}_{A_0}(\mathcal{D}),\ D^{\geq 0}_{A_0}(\mathcal{D}))$ *via the equivalence* F_0 (*see the end of Section* 3.2.1 *for notation*).

Corollary 3.9. *The* t*-structure on* $\mathrm{Coh}^f(\widetilde{\mathcal{N}})$ *induced by the exotic* t*-structure coincides with the one induced from* $(D^{<0}_{A_0}(\mathcal{D}),\ D^{\geq 0}_{A_0}(\mathcal{D}))$ *via the equivalence* F_0 (*see the end of* 3.2.1).

In particular, we have a Morita equivalence $\mathcal{A}_e \sim U\text{-mod}^{\lambda}_{\hat{e}}$ *for every* p*-regular* $\lambda \in \Lambda$.

The corollary follows by comparing Theorem 3.8 with Theorem 2.13.

Corollary 3.10. (a) *Let* U^{λ}_e *denote the specialization of the enveloping algebra* $U(\mathfrak{g})$ *at the central character corresponding to* $e \in \mathcal{N}$ *and a regular integral weight* λ. *Then we have a canonical isomorphism* $K(U^{\lambda}_e)_F \cong H^{\mathrm{BM}}_{\bullet}(\mathcal{B}_e)_F$, *where* F *is a field of characteristic zero.*

(b) *The image of the set of classes of irreducible modules under this isomorphism is independent of the base field* $\boldsymbol{k}$, *except for a finite number of values of the characteristic.*

Part (a) of the corollary follows directly from Theorem 3.5 (cf. the discussion preceding Proposition 2.16), while part (b) follows from Corollary 3.9 and Proposition 2.27.

Remark 3.11. For $e = 0$ part (a) of the proposition is standard, and part (b) can be deduced from [1]. Our method uses the principal tool of [1], namely, the reflection functors, in the disguise of the braid group action; geometry of the Springer map is the new ingredient.

In fact, we have the following stronger, though more difficult statement. We will say that a basis in $H^{\mathrm{BM}}_{\bullet}(\mathcal{B}_e)$ is canonical if it is the image of a basis in the equivariant Grothendieck group $K^{\mathbb{G}_m}(\mathcal{B}_e)$, satisfying Lusztig's axioms [45], under forgetting the equivariance composed with the Chern character map. According to a result of [45] such a basis is unique up to multiplication of some of its elements by -1, if it exists.

Corollary 3.12. *Enforce the assumption of Corollary* 2.23. *Then for almost all* $p = \mathrm{char}(\boldsymbol{k})$ *the isomorphism* $K^0(U\text{-mod}^0_e)_F \cong H^{\mathrm{BM}}_*(\mathcal{B}_e)_F$ *of Corollary* 3.10 (a) *sends classes of irreducible objects to elements of a canonical basis. Thus Conjecture* 17.2 *of* [45] *holds in this case.*

This corollary is immediate from Corollary 2.23 together with Theorem 3.8(d). Thus its proof, unlike the proof of Corollary 3.10, relies on Gabber's Theorem [7] and ideas of local geometric Langlands duality, on which the results of [2] are based.

Remark 3.13. For large p the particular case $e = 0$ of the Conjecture 17.2 of [45] is well known to imply the previous Lusztig conjectures [46], which describe characters of algebraic groups in finite characteristics. Lusztig's program for a proof of these

conjectures for indefinitely large p has been carried out by several authors. An alternative proof is given in [3].

None of the available methods gives a proof of the conjectures for some particular fixed value of p.

Notice that the strategy of proof of this conjecture (for indefinitely large p) outlined above does not use quantum groups.

4. Perverse sheaves on affine flags of the dual group (local geometric Langlands)

4.1. Generalities on geometric Langlands duality. Recall that ${}^LG$ is the group dual to G in the sense of Langlands. Several good surveys of geometric Langlands duality program has appeared recently [33], [34], [37], so I will only briefly recall the set-up.

The geometric Langlands duality is a categorification of the classical Langlands duality for function fields. The latter seeks to attach an automorphic form to a homomorphism from a version of the Galois group to ${}^LG$. In other words, the problem is to provide a spectral decomposition for Hecke operators acting in the space of automorphic functions, and relate the space of spectral parameters to homomorphisms of the Galois group to the dual group. As was probably first observed by A. Weil, in the case of a function field the automorphic space in question is the set of isomorphism classes of G-bundles, possibly with an additional level structure, on an algebraic curve over $\mathbb{F}_q$. Thus it is the set of $\mathbb{F}_q$ points of the corresponding moduli space (stack).

Passage to the geometric duality theory is based on the following variation of Grothendieck's sheaf-function correspondence principle. The variation says that for an algebraic variety (or stack) over $\mathbb{F}_q$ a natural categorification of the space of functions on the set $X(\mathbb{F}_q)$ is the derived category of l-adic sheaves on X. Thus the objective of the geometric duality theory is a spectral decomposition of the derived category of l-adic sheaves on a moduli space of G-bundles, where the space of spectral parameters is identified with the space of ${}^LG$ local systems. It is a non-trivial, and not completely solved, problem to assign a formal meaning to the previous sentence; however, in some cases it amounts to an equivalence between the l-adic derived category of the moduli stack and the derived category of coherent sheaves on a stack mapping to the stack of local systems.

The above formulations referred to a more developed *global* version of the theory. However, the classical Langlands conjectures have both a global and a local version. The global one provides a conjectural classification of automorphic representations of the group of adele points of a reductive group over a global field, i.e. either a number field, or the field of rational functions on a curve over a finite field. The local one describes all irreducible representations of a reductive group over a local field; recall that a local function field is a field of formal Laurent series $\mathbb{F}_q((t))$. The geometric theory studies the derived category of l-adic sheaves on homogeneous spaces of the

formal loop group $\boldsymbol{^L G}((t))$ of the dual group ${}^L G$. It is a group ind-scheme over $\mathbb{F}_q$, whose group of $\mathbb{F}_q$ points is identified with ${}^L G(\mathbb{F}_q((t)))$. The results are expected to link such l-adic derived categories to coherent sheaves on spaces related to G-local systems on the punctured formal disc, cf. [35].

4.2. Results of [2], [12], [15]. Some particular results of this type have been achieved in *loc. cit.*

4.2.1. Statement of a result. Recall that the Iwahori subgroup $I \subset {}^L G((\mathbb{F}_q(t)))$ consists of those maps from the punctured formal disc to ${}^L G$, which can be extended to a map from the whole disc, so that the image of the closed point is contained in a fixed Borel subgroup ${}^L B \subset {}^L G$.

We have a group subscheme (pro-algebraic group) $\boldsymbol{I} \subset \boldsymbol{^L G}((t))$, such that $\boldsymbol{I}(\mathbb{F}_q) = I$. The affine flag space $\mathcal{F}\ell$ of ${}^L G$ is the homogeneous space $\boldsymbol{^L G}((t))/\boldsymbol{I}$. It is an ind-algebraic variety such that $\mathcal{F}\ell(\mathbb{F}_q) = {}^L G(\mathbb{F}_q((t)))/I$. The group $\boldsymbol{I}$ acts on $\mathcal{F}\ell$. The orbits of this action, called affine Schubert cells, are in bijection with the affine Weyl group W_{aff}.

Let $\mathcal{P}$ denote the category of perverse sheaves on $\mathcal{F}\ell$, which are equivariant with respect to the prounipotent radical of $\boldsymbol{I}$. Let $\mathcal{P}^I \subset \mathcal{P}$ be the full subcategory of $\boldsymbol{I}$ equivariant sheaves.

Let ${}^{nf}\mathcal{P}^I \subset \mathcal{P}^I$ be the Serre subcategory generated by irreducible objects, corresponding to those $w \in W_{\mathrm{aff}}$, which are not the minimal length representatives of a left W coset. Let ${}^f\mathcal{P} = \mathcal{P}/{}^{nf}\mathcal{P}$ be the Serre quotient category.

Remark 4.1. To clarify the definition of ${}^f\mathcal{P}$ we remark that this category can be also described as the category of Iwahori–Whittaker sheaves [2]. Thus it is related to the Whittaker model, which is one of the main tools in representation theory of reductive groups over local and global fields.

Theorem 4.2. (a) ([15]) *We have a canonical equivalence* $D^b(\mathcal{P}) \cong D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathcal{N}})$.

(b) ([2]) *We have a canonical equivalence* $D^b({}^f\mathcal{P}) \cong D^G(\widetilde{\mathcal{N}})$. *The image of* ${}^f\mathcal{P}$ *under this equivalence consists of equivariant exotic sheaves.*

The theorem is motivated by the known isomorphisms of Grothendieck groups; the question of possibility of such (or similar) equivalence has been raised, e.g., by V. Ginzburg, see Introduction to [31]. More precisely, the Grothendieck groups of the two categories appearing in Theorem 4.2 (a) are isomorphic to the group algebra of the affine Weyl group of ${}^L G$. A more interesting version of the isomorphism is obtained by replacing the categories by their graded version: $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathcal{N}})$ is replaced by $D^{G \times \mathbb{G}_m}(\widetilde{\mathfrak{g}} \times \widetilde{\mathcal{N}})$, while the definition of the graded version of $\mathcal{P}$ is more subtle (cf. Proposition 4.5 below and also [9]). The corresponding Grothendieck groups turn out to be isomorphic to *the affine Hecke algebra*, see [31], [44].

Similarly, the Grothendieck groups of both categories appearing in Theorem 4.2 (b) are identified with the *anti-spherical* module over the extended affine Weyl group,

while in the graded version of the theory we get the anti-spherical module over the affine Hecke algebra. Here by the anti-spherical module we mean the induction of the sign representation from the finite Weyl group (respectively, Hecke algebra) to the affine one.

I would like to emphasize that this isomorphism of the two realizations of the affine Hecke algebra is the key step in the proof of classification of its representations due to Kazhdan and Lusztig [43] (see also [31]), which establishes a particular case of the local Langlands conjecture. This is another illustration of the relation of Theorem 4.2 to local Langlands duality.

The proof of Theorem 4.2 builds on previously known constructions of categories related to G in terms of perverse sheaves on homogeneous spaces for ${}^{\boldsymbol{L}}\boldsymbol{G}((t))$. The first important result is *the geometric Satake isomorphism* [39], [47], [8], which identifies the tensor category $\mathrm{Rep}(G)$ of algebraic representations with the category of perverse sheaves on the affine Grassmannian $\mathcal{G}r = {}^{\boldsymbol{L}}\boldsymbol{G}((t))/{}^{L}\boldsymbol{G}_O$ equivariant with respect to ${}^{L}\boldsymbol{G}_O$. Here ${}^{L}\boldsymbol{G}_O \subset {}^{\boldsymbol{L}}\boldsymbol{G}((t))$ is the group subscheme, such that ${}^{L}\boldsymbol{G}_O(\mathbb{F}_q)$ consists of maps which extend to the non-punctured disc. Furthermore, Gaitsgory [36] used this result to provide a categorification of the description of the *center* of the affine Hecke algebra. Using some ideas of I. Mirković we observe that the so-called Wakimoto sheaves provide a categorification of the maximal abelian subalgebra in the affine Hecke algebra due to Bernstein, see, e.g., [31], [44]. The maximal projective object in the category of sheaves on the finite dimensional flag variety of ${}^{L}G$ smooth along the Schubert stratification, which plays a central role in Soergel's description of category O, cf. [50], is a categorification of the q anti-symmetrizer (an element of the finite Hecke algebra, which acts by zero in all irreducible representation except for the sign representation). Under the equivalence of Theorem 4.2 (a) it corresponds to the structure sheaf of $\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathcal{N}}$. A combination of these ingredients yields a proof of the theorem.

4.2.2. Possible generalizations. It is natural to ask if the multiplication in the affine Hecke algebra corresponds to a monoidal structure on the derived categories of coherent sheaves and constructible sheaves appearing in Theorem 4.2 (a). In order to get such a monoidal structure, we need to replace the categories defined above by closely related ones with the same Grothendieck group. One way to do it is as follows. Let $\boldsymbol{I}'$ be the pro-unipotent radical of $\boldsymbol{I}$. Let $\mathcal{P}'$ be the category of perverse sheaves on "the basic affine space" ${}^{\boldsymbol{L}}\boldsymbol{G}((t))/\boldsymbol{I}'$, which are $\boldsymbol{I}$-monodromic with unipotent monodromy. Then convolution provides the derived category $D^b(\mathcal{P}')$ with a monoidal structure. Notice that this monoidal category does not have a unit object, though this can be repaired by adding some pro-objects to the category, the unit object is then the free pro-unipotent local system on $\boldsymbol{I}/\boldsymbol{I}' \subset {}^{\boldsymbol{L}}\boldsymbol{G}((t))/\boldsymbol{I}'$.

On the dual side we consider the category $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})$. One can show that convolution provides this category with a monoidal structure. Let $\mathrm{Coh}^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})' \subset \mathrm{Coh}^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})$ denote the full subcategory of complexes, whose cohomology sheaves are set-theoretically supported on the preimage of $\mathcal{N} \subset \mathfrak{g}$. A standard

argument shows that it yields a full embedding of derived categories $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})' := D^b(\mathrm{Coh}^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})')$ into $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})$. The full subcategory $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})' \subset D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})$ is closed under the convolution product, though it does not contain the unit object $\delta_*(\mathcal{O})$, where $\delta\colon \widetilde{\mathfrak{g}} \to \widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}}$ is the diagonal embedding.

It is easy to see that the push-forward (respectively, pull-back) functors $\mathrm{Coh}^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathcal{N}}) \to \mathrm{Coh}^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})'$, $\mathcal{P} \to \mathcal{P}'$ induce isomorphisms of Grothendieck groups.

Theorem 4.3 ([15]). *We have a natural monoidal equivalence* $D(\mathcal{P}') \cong D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})'$.

Remark 4.4. Another version of Theorem 4.2 links the monoidal $\boldsymbol{I}$ equivariant derived category to the monoidal derived category of G-equivariant coherent sheaves on the fiber square of $\widetilde{\mathcal{N}}$ over $\mathfrak{g}$. An additional subtlety in this case is related to nonvanishing of $\mathrm{Tor}^{\mathcal{O}_{\mathfrak{g}}}_{>0}(\mathcal{O}_{\widetilde{\mathcal{N}}}, \mathcal{O}_{\widetilde{\mathcal{N}}})$. One actually has to take these Tor groups into account by working with the *derived fiber product*, which is a differential-graded scheme, rather than an ordinary scheme. This issue does not arise in the other settings mentioned above, because $\mathrm{Tor}^{\mathcal{O}_{\mathfrak{g}}}_{>0}(\mathcal{O}_{\widetilde{\mathfrak{g}}}, \mathcal{O}_{\widetilde{\mathfrak{g}}}) = 0 = \mathrm{Tor}^{\mathcal{O}_{\mathfrak{g}}}_{>0}(\mathcal{O}_{\widetilde{\mathfrak{g}}}, \mathcal{O}_{\widetilde{\mathcal{N}}})$. However, one has to work with differential graded schemes in order to define the convolution product on $D^G(\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathfrak{g}})$.

4.2.3. Relation to the material of Section 2. Many of the constructions from Section 2 are motivated by the equivalences of Theorem 4.2.

For example, the categories $D(\mathcal{P})$, $D(\mathcal{P}')$ carry a natural B_{aff} action by *Radon transforms*, cf. beginning of Section 3.2.2, where a similar structure for a finite dimensional flag variety is mentioned. To define the action we recall that the $\boldsymbol{L}\boldsymbol{G}((t))$ orbits on $\mathcal{F}\ell^2$ are indexed by the affine Weyl group. If $\mathcal{F}\ell^2_w$ is the orbit corresponding to $w \in W_{\mathrm{aff}}$, and $\mathrm{pr}^w_i\colon \mathcal{F}\ell^2_w \to \mathcal{F}\ell$ are the projections, where $i = 1, 2$ then we define a functor $R_w\colon D(\mathcal{P}) \to D(\mathcal{P})$ by $R_w(\mathcal{F}) = \mathrm{pr}^w_{2*}\mathrm{pr}^{w*}_1(\mathcal{F})$. Then we have an action of B_{aff} on $D(\mathcal{P})$, $D({}^f\mathcal{P})$, such that $\widetilde{w} \mapsto R_w$. Under the equivalences of Theorem 4.2 (b) this action corresponds to the action described in Section 2.

Finally, I would like to quote the statement that allows to link the grading on Ext spaces appearing in Theorem 2.19 to Frobenius weights, thus providing a way to prove Theorem 2.19. To state it we introduce the following notation. Let Φ be either of the two equivalences appearing in Theorem 4.2. Let Fr be the autoequivalence of the corresponding category of constructible sheaf, sending a sheaf to its pull-back under the Frobenius morphism. Let $\boldsymbol{q}$ be an automorphism of either $\widetilde{\mathcal{N}}$ or $\widetilde{\mathfrak{g}} \times_{\mathfrak{g}} \widetilde{\mathcal{N}}$ given by $(\mathfrak{b}, x) \mapsto (\mathfrak{b}, qx)$ or $(\mathfrak{b}_1, \mathfrak{b}_2, x) \mapsto (\mathfrak{b}_1, \mathfrak{b}_2, qx)$ respectively; here q stands for the cardinality of the base finite field $\mathbb{F}_q$.

Proposition 4.5 (cf. [2]). *We have a canonical isomorphism* $\Phi \circ \boldsymbol{q}^* \cong \mathrm{Fr} \circ \Phi$.

References

[1] Andersen, H. H., Jantzen, J. C., Soergel, W., Representations of quantum groups at a pth root of unity and of semisimple groups in characteristic p: independence of p. *Astérisque* **220** (1994), 321 pp.

[2] Arkhipov, S., Bezrukavnikov, R., Perverse sheaves on affine flags and Langlands dual group. *Israel Math. J.*, to appear; math.RT/0201073.

[3] Arkhipov, S., Bezrukavnikov, R., Ginzburg, V., Quantum Groups, the loop Grassmannian, and the Springer resolution. *J. Amer. Math. Soc.* **17** (3) (2004), 595–678

[4] Backelin, E., Kremnitzer, K., Localization for quantum groups at a root of unity. Preprint; math.RT/0407048.

[5] Beilinson, A., Bernstein, J., Localisation de g-modules. *C. R. Acad. Sci. Paris Sér. I Math.* **292** (1) (1981), 15–18.

[6] Beilinson, A., Bernstein, J., A generalization of Casselman's submodule theorem. In *Representation theory of reductive groups* (Park City, Utah, 1982), Progr. Math. 40, Birkhäuser, Boston, MA, 1983, 35–52.

[7] Beilinson, A., Bernstein, J., Deligne, P., Faisceaux pervers. *Astérisque* **100** (1982), 5–171.

[8] Beilinson, A., Drinfeld, V., Quantization of Hitchin's Integrable System and Hecke Eigensheaves. Preprint; http://www.math.uchicago.edu/~arinkin/langlands/

[9] Beilinson, A., Ginzburg, V., Soergel, W., Koszul duality patterns in representation theory. *J. Amer. Math. Soc.* **9** (2) (1996), 473–527.

[10] Belov-Kanel, A., Kontsevich, M., Automorphisms of the Weyl algebra. *Lett. Math. Phys.* **74** (2) (2005), 181–199.

[11] van den Bergh, M., Noncommutative crepant resolutions. In *The legacy of Niels Henrik Abel*, Springer-Verlag, Berlin 2004, 749–770.

[12] Bezrukavnikov, R., Perverse sheaves on affine flags and nilpotent cone of the Langlands dual group. *Israel J. Math.*, to appear.

[13] Bezrukavnikov, R., Perverse coherent sheaves (after Deligne). Preprint; math.AG/0005152.

[14] Bezrukavnikov, R., Cohomology of tilting modules over quantum groups and t-structures on derived categories of coherent sheaves. *Invent. Math.*, to appear.

[15] Bezrukavnikov, R., An equivalence between two categorical realization of the affine Hecke algebra. In preparation.

[16] Bezrukavnikov, R., Finkelberg, M., Ginzburg, V., Cherednik algebras and Hilbert schemes in characteristic p. *Represent. Theory* **10** (2006), 254–298.

[17] Bezrukavnikov, R., Kaledin, D., McKay equivalence for symplectic resolutions of quotient singularities. *Tr. Mat. Inst. Steklova* **246** (2004), Algebr. Geom. Metody, Svyazi i Prilozh., 20–42; English transl. *Proc. Steklov Inst. Math.* **246** (3) (2004), 13–33.

[18] Bezrukavnikov, R., Kaledin, D., Fedosov quantization in positive characteristic. Preprint; math.AG/0501247.

[19] Bezrukavnikov, R., Mirkovic, I., Rumynin, D., Localization of modules for a semisimple Lie algebra in prime characteristic. *Ann. of Math.*, to appear; math.RT/0205144.

[20] R. Bezrukavnikov, Mirkovic, I., Rumynin, D., Singular localization and intertwining functors for semisimple Lie algebras in prime characteristic. *Nagoya Math. J.*, submitted; math.RT/0602075

[21] Bondal, A., Kapranov, M., *Representable functors, Serre functors, and mutations. Izv. Akad. Nauk* **35** (3) (1990), 519–541.

[22] Bondal, A., Orlov, D., Derived categories of coherent sheaves. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 47–56.

[23] Bridgeland, T., Derived categories of coherent sheaves. *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 563–582.

[24] Bridgeland, T., Stabilities on triangulated categories. *Ann. of Math.*, to appear; math.AG/0212237.

[25] Bridgeland, T., Stability conditions on K3 surfaces. Preprint; math.AG/0307164.

[26] Bridgeland, T., Stability conditions on a non-compact Calabi-Yau threefold. Preprint; math.AG/0509048.

[27] Bridgeland, T., Stability conditions and Kleinian singularities. Preprint; math.AG/0508257.

[28] Bridgeland, T., Flops and derived categories. *Invent. Math.* **147** (3) (2002), 613–632.

[29] Brown, K., Gordon, I., The ramification of centres: Lie algebras in positive characteristic and quantised enveloping algebras. *Math. Z.* **238** (4) (2001), 733–779.

[30] Brylinski, J.-L., Kashiwara, M., Kazhdan-Lusztig conjecture and holonomic systems. *Invent. Math.* **64** (3) (1981), 387–410.

[31] Chriss, N., Ginzburg, V., *Representation theory and complex geometry.* Birkhäuser, Boston, MA, 1997.

[32] De Concini, C., Kac, V., Representations of quantum groups at roots of 1. In *Operator algebras, unitary representations, enveloping algebras, and invariant theory* (Paris, 1989), Progr. Math. 92, Birkhäuser, Boston, MA, 1990, 471–506.

[33] Frenkel, E., Recent advances in the Langlands program. *Bull. Amer. Math. Soc.* (*N.S.*) **41** (2) (2004), 151–184.

[34] Frenkel, E., Lectures on the Langlands Program and Conformal Field Theory. Preprint; hep-th/0512172.

[35] Frenkel, E., Gaitsgory, D., Local geometric Langlands correspondence and affine Kac-Moody algebras. Preprint; math.RT/0508382.

[36] Gaitsgory, D., Construction of central elements in the affine Hecke algebra via nearby cycles. *Invent. Math.* **144** (2) (2001), 253–280.

[37] Gaitsgory, D., Geometric Langlands correspondence for GL_n. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 571–582.

[38] Gan, W.-L., Ginzburg, V., Quantization of Slodowy slices. *Internat. Math. Res. Notices* **5** (2002), 243–255.

[39] Ginzburg, V., Perverse sheaves on a Loop group and Langlands' duality. Preprint; alg-geom/9511007.

[40] Hürlimann, W., Sur le groupe de Brauer d'un anneau de polynômes en caractéristique p et la théorie des invariants. In *The Brauer group* (Les Plans-sur-Bex, 1980), Lecture Notes in Math. 844, Springer, Berlin 1981, 229–274.

[41] Kaledin, D., Derived equivalences by quantization. Preprint; math.AG/0504584.

[42] Kashiwara, M., On crystal bases of the q-analogue of universal enveloping algebras. *Duke Math. J.* **63** (1991), 465–516.

[43] Kazhdan, D., Lusztig, G., Proof of the Deligne-Langlands conjecture for Hecke algebras. *Invent. Math.* **87** (1) (1987), 153–215.

[44] Lusztig, G., Bases in equivariant K-theory, *Represent. Theory* **2** (1998), 298–369.

[45] Lusztig, G., Bases in equivariant K-theory II, *Represent. Theory* **3** (1999), 281–353.

[46] Lusztig, G., Some problems in the representation theory of finite Chevalley groups. In *The Santa Cruz Conference on Finite Groups* (Univ. California, Santa Cruz, Calif., 1979), Proc. Symp. Pure Math. 37, Amer. Math. Soc., Providence, RI, 1980, 313–317.

[47] Mirković, I., Vilonen, K., Geometric Langlands duality and representations of algebraic groups over commutative rings. *Ann. of Math.*, to appear; math.RT/0401222.

[48] Ogus, A., Vologodsky, V., Nonabelian Hodge Theory in Characteristic p. Preprint: math.AG/0507476.

[49] Premet, A., Special transverse slices and their enveloping algebras (With an appendix by Serge Skryabin). Adv. Math. **170** (1) (2002), 1–55.

[50] Soergel, W., Gradings on representation categories. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 800–806.

Department of Mathematics, MIT, Cambridge, MA 02139, U.S.A.

E-mail: bezrukav@math.mit.edu

Spaces of quasi-maps into the flag varieties and their applications

Alexander Braverman

Abstract. Given a projective variety X and a smooth projective curve C one may consider the moduli space of maps $C \to X$. This space admits certain compactification whose points are called quasi-maps. In the last decade it has been discovered that in the case when X is a (partial) flag variety of a semi-simple algebraic group G (or, more generally, of any symmetrizable Kac–Moody Lie algebra) these compactifications play an important role in such fields as geometric representation theory, geometric Langlands correspondence, geometry and topology of moduli spaces of G-bundles on algebraic surfaces, 4-dimensional super-symmetric gauge theory (and probably many others). This paper is a survey of the recent results about quasi-maps as well as their applications in different branches of representation theory and algebraic geometry.

Mathematics Subject Classification (2000). Primary 22E46; Secondary 14J60, 14J81.

Keywords. Quasi-maps, Schubert varieties, geometric Langlands duality, supersymmetric gauge theory.

1. Introduction

The spaces of quasi-maps into the flag varieties were introduced by V. Drinfeld about 10 years ago and since then proved to play an important role in various parts of geometric representation theory; more recently it was discovered that some related constructions are useful also in more classical algebraic geometry as well as in some questions coming from mathematical physics.

This paper constitutes at attempt to give a more or less self-contained presentation of the results related to such spaces. The origin of quasi-maps is as follows: let C be a smooth projective algebraic curve (over an algebraically closed field k) and let $X \subset \mathbb{P}^N$ be a projective variety over k. One can look at the space $\mathrm{Maps}^d(C, X)$ of maps $C \to X$ such that the composite map $C \to X \to \mathbb{P}^N$ has degree $d \in \mathbb{Z}_+$. These are quasi-projective schemes of finite type; in many problems of both representation theory and algebraic geometry it is important to have a natural compactification of this scheme; one compactification of this sort is provided by the space $\mathrm{QMaps}^d(C, X)$ of *quasi-maps* from C to X (cf. Section 2 for the precise definition). The main property

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

of the scheme $\mathrm{QMaps}^d(C, X)$ is that it possesses a stratification of the form

$$\mathrm{QMaps}^d(C, X) = \bigcup_{d'=0}^{d} \mathrm{Maps}^{d'}(C, X) \times \mathrm{Sym}^{d-d'}(C)$$

where $\mathrm{Sym}^a(C)$ denote the a-th symmetric power of the curve C. In other words, in order to specify a point of $\mathrm{QMaps}^d(C, X)$ one must specify an honest map $C \to X$ of degree $d' \leq d$ together with $d - d'$ unordered points of C.

We must warn the reader from the very beginning that the scheme $\mathrm{QMaps}^d(C, X)$ depends on the embedding $X \subset \mathbb{P}^N$. However, in many cases such an embedding is given to us in the original problem. More generally, when X is a closed subscheme of a product $\mathbb{P}^{N_1} \times \cdots \times \mathbb{P}^{N_l}$ we may speak about $\mathrm{QMaps}^{d_1,\ldots,d_l}(C, X)$. For example, if X is the complete flag variety of a semi-simple algebraic group G then X has a canonical embedding as above (in this case l is the rank of G). We discuss the details in Section 2.

We then turn to applications of quasi-maps. In Section 3 we explain the relation between quasi-maps spaces and the so-called *semi-infinite Schubert varieties*. In particular, we explain the calculation of the Intersection Cohomology sheaf of the quasi-maps spaces and relate it to Lusztig's periodic polynomials. We also mention that quasi-maps could be used to construct some version of the category of perverse sheaves on the (still not rigorously defined) *semi-infinite flag variety* and relate this category with the category of representations of the so-called *small quantum group*.

In Section 4 we discuss the results of [10] where the stacks $\overline{\mathrm{Bun}}_B$ are used in order to construct the so-called *geometric Eisenstein series* (thus the contents of [10] have to do with application of the stacks $\overline{\mathrm{Bun}}_B$ (which are close relatives of the scheme $\mathrm{QMaps}(C, X)$) to geometric Langands correspondence; the rest of the paper is independent of this section and therefore and can be easily skipped by a non-interested reader).

In Section 5 we discuss quasi-maps into affine (partial) flag varieties and their relation to the Uhlenbeck compactifications of moduli spaces of G-bundles on algebraic surfaces. In Section 6 we explain how to apply these constructions to certain enumerative questions related to quantum cohomology of the flag manifolds as well as to $N = 2$ super-symmetric 4-dimensional gauge theory. Section 7 is devoted to the discussion of some open questions related to the above subjects.

Acknowledgements. This paper is mostly based on the author's joint papers with various people including S. Arkhipov, R. Bezrukavnikov, P. Etingof, M. Finkelberg, D. Gaitsgory and I. Mirkovic; I am grateful to all of them for being very fruitful and patient collaborators. I am also grateful to J. Bernstein, V. Drinfeld and D. Kazhdan for their constant guidance and to H. Nakajima, N. Nekrasov and A. Okounkov for interesting and illuminating discussions related to Section 6 of this paper.

2. Definition of quasi-maps

In this section we introduce quasi-maps' spaces and some of their relatives. The reader may skip the details for most applications.

2.1. Maps and quasi-maps into a projective variety. Let X be a closed subvariety of the projective space $\mathbb{P}^N$ and let C be a smooth projective curve. For any integer $d \geq 0$ we may consider the space $\mathrm{Maps}^d(C, X)$ consisting of maps $C \to X$ such that the composition $C \to X \to \mathbb{P}^N$ has degree d. This space has a natural scheme structure and it is in fact quasi-projective. However, it is well known that in general it does not have to projective (in fact it is almost never projective).

Example. Let $X = \mathbb{P}^N$. In this case $\mathrm{Maps}^d(C, X)$ classifies the following data:

- A line bundle $\mathcal{L}$ on C of degree $-d$.
- An embedding of vector bundles $\mathcal{L} \hookrightarrow \mathcal{O}_C^{N+1}$.

The reason is that every such embedding defines a one-dimensional subspace in $\mathbb{C}^{N+1}$ for every point $c \in C$ and thus we get a map $C \to \mathbb{P}^N$.

Consider, for example, the case when $C = \mathbb{P}^1$. In that case $\mathcal{L}$ must be isomorphic to the line bundle $\mathcal{O}_{\mathbb{P}^1}(-d)$ (note that such an isomorphism is defined uniquely up to a scalar) and thus $\mathrm{Maps}^d(\mathbb{P}^1, \mathbb{P}^N)$ becomes an open subset in the projectivization of the vector space $\mathrm{Hom}(\mathcal{O}_{\mathbb{P}^1}(-d), \mathcal{O}_{\mathbb{P}^1}^{N+1}) \simeq \mathbb{C}^{(N+1)(d+1)}$, i.e. $\mathrm{Maps}^d(\mathbb{P}^1, \mathbb{P}^N)$ is an open subset of $\mathbb{P}^{(N+1)(d+1)-1}$. The reason that it does not coincide with it is that not every non-zero map $\mathcal{O}_{\mathbb{P}^1}(-d) \to \mathcal{O}_{\mathbb{P}^1}^{N+1}$ gives rise a map $\mathbb{P}^1 \to \mathbb{P}^N$ – we need to consider only those maps which do not vanish in every fiber.

The above example suggests the following compactification of $\mathrm{Maps}^d(C, X)$. Namely, we define the space of *quasi-maps from C to X of degree d* (denoted by $\mathrm{QMaps}^d(C, X)$) to be the scheme classifying the following data:

1) A line bundle $\mathcal{L}$ on C of degree $-d$.

2) A non-zero map $\kappa\colon \mathcal{L} \to \mathcal{O}_C^{N+1}$.

3) Note that κ defines an honest map $U \to \mathbb{P}^N$ where U is an open subset of C. We require that the image of this map lies in X.

For example it is easy to see that if $X = \mathbb{P}^N$ and $C = \mathbb{P}^1$ then $\mathrm{QMaps}^d(C, X) \simeq \mathbb{P}^{(N+1)(d+1)-1}$.

In general $\mathrm{QMaps}^d(C, X)$ is projective. Also, set-theoretically it can be explicitly described in the following way. Assume that we are given a quasi-map $(\mathcal{L}, \kappa)$ as above. Then κ might have zeros at points $c_1, \ldots, c_k$ of C of order $a_1, \ldots, a_k$ respectively. On the other hand, it follows from 3) above that κ defines an honest map from the complement to the points $c_1, \ldots, c_k$ to X. Since X is projective this map can be extended to the whole of C. Let us call this map κ'. It is easy to see that κ' has degree $d - \sum a_i$. Also one can recover κ from κ' and the collection $(c_1, a_1), \ldots, (c_k, a_k)$. Thus it follows that $\mathrm{QMaps}^d(C, X)$ is equal to the disjoint union of locally closed

subvarieties of the following form:

$$\mathrm{QMaps}^d(C, X) = \bigcup_{0 \leq d' \leq d} \mathrm{Maps}^{d'}(C, X) \times \mathrm{Sym}^{d-d'}(C). \tag{2.1}$$

Here $\mathrm{Sym}^{d-d'}(C)$ denotes the corresponding symmetric power of C.

Here is a generalization of the above construction. Assume that X is embedded into a product $\mathbb{P}^{N_1} \times \cdots \times \mathbb{P}^{N_k}$ of projective spaces. Then, in a similar fashion one can talk about $\mathrm{Maps}^{d_1,\ldots,d_l}(C, X)$ (here all $d_i \geq 0$) and $\mathrm{QMaps}^{d_1,\ldots,d_l}(C, X)$.

2.2. The case of complete flag varieties. Let now G be a semi-simple simply connected algebraic group over k and let $\mathfrak{g}$ denote its Lie algebra. We want to take X to be the complete flag variety of G. If we choose a Borel subgroup B of G then $X = G/B$. We shall sometimes denote this variety by $X_{G,B}$ (later we shall also consider the *partial flag varieties* G/P associated with a parabolic subgroup $P \subset G$; this variety will be denoted by $X_{G,P}$).

Let $V_1, \ldots, V_l$ denote the fundamental representations of G. It is well known that $X_{G,B}$ has a canonical (Plücker) embedding into $\prod_{i=1}^{l} \mathbb{P}(V_i^*)$. This enables us to talk about quasi-maps into X.

We can describe the set of parameters $(d_1, \ldots, d_l)$ in a little bit more invariant terms. First, let us denote by T the Cartan group of G and let Λ_G denote the coweight lattice of G; by definition $\Lambda_G = \mathrm{Hom}(\mathbb{C}^*, T)$ (in the case $k = \mathbb{C}$). We have the natural well-known identification $\Lambda_G = H_2(X, \mathbb{Z})$. This allows us to talk about maps $C \to X$ of degree $\theta \in \Lambda_G$. Also if we let $(\omega_1, \ldots, \omega_l)$ denote the fundamental weights of G then we can also identify Λ_G with $\mathbb{Z}^l$ by sending a coweight θ to $d_1 = (\theta, \omega_1), \ldots d_l = (\lambda, \theta_l)$. Under these identification $\mathrm{Maps}^{\theta}(C, X)$ is the same as $\mathrm{Maps}^{d_1,\ldots,d_l}(C, X)$ in the sense of the previous subsection. It is also clear that this space may be non-empty only if all $d_i \geq 0$. We say that θ is positive if all $d_i \geq 0$ and denote the semigroup of all positive θ's by Λ_G^+.

In the case $C = \mathbb{P}^1$ we shall denote the space $\mathrm{Maps}^{\theta}(C, X_{G,B})$ by $\mathcal{M}^{\theta}_{G,B}$ and the space $\mathrm{QMaps}^{\theta}(C, X_{G,B})$ by $\mathcal{QM}^{\theta}_{G,B}$.

2.3. Laumon's resolution. Consider the case $G = \mathrm{SL}(n)$ (thus $l = n - 1$). In this case $X_{G,B}$ is just the variety of complete flags $0 \subset V_1 \subset V_2 \subset \cdots \subset V_n = \mathbb{C}^n$, $\dim V_i = i$. Thus, a map $C \to X_{G,B}$ is the same as a complete flag of subbundles

$$0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n = \mathcal{O}_C^n.$$

where the rank of $\mathcal{V}_i$ is equal to i. Also we have $d_i = -\deg \mathcal{V}_i$.

Define now the space $\mathrm{QMaps}^{L,\theta}(C, X_{G,B})$ to consist of all flags as above where $\mathcal{V}_i$ is an arbitrary *subsheaf* of $\mathcal{O}_C^n$ of degree $-d_i$. The space was considered by G. Laumon in [32]. It is known (cf. [30]) that the natural open embedding of $\mathrm{Maps}^{\theta}(C, X_{G,B})$ into both $\mathrm{QMaps}^{\theta}(C, X_{G,B})$ and $\mathrm{QMaps}^{L,\theta}(C, X_{G,B})$ extends to a projective morphism $\mathrm{QMaps}^{L,\theta}(C, X_{G,B}) \to \mathrm{QMaps}^{\theta}(C, X_{G,B})$. In the case $C = \mathbb{P}^1$ the space

$\mathrm{QMaps}^{L,\theta}(\mathbb{P}^1, X_{G,B})$ is smooth and provides in fact a small resolution of singularities of $\mathrm{QMaps}^{\theta}(\mathbb{P}^1, X_{G,B})$.

2.4. The stacks $\overline{\mathbf{Bun}}_{\boldsymbol{B}}$. Let us fix a curve C as above and let G again be a semi-simple simply connected algebraic group with a Borel subgroup B (more generally, one can assume that G is any reductive group whose derived group is simply connected; e.g. one may also consider the case $G = \mathrm{GL}(n)$). We may consider the algebraic stack $\mathrm{Bun}_G = \mathrm{Bun}_G(C)$ classifying principal algebraic G-bundles on C. Similarly we may consider the stack Bun_B which classifies B-bundles. The embedding $B \to G$ gives rise to a the natural morphism $p\colon \mathrm{Bun}_B \to \mathrm{Bun}_G$. In the case $G = \mathrm{GL}(n)$ the stack Bun_G classifies vector bundles of rank n on C and the stack Bun_B classifies flags of the forms

$$0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n$$

where each $\mathcal{V}_i$ is a vector bundle of rank i on C and the embedding $\mathcal{V}_i \to \mathcal{V}_{i+1}$ are embeddings of vector bundles.

We have the natural projection $B \to T$ (where T as before denotes the Cartan group of G). Hence we also have the natural map $q\colon \mathrm{Bun}_B \to \mathrm{Bun}_T$. In the case $G = \mathrm{GL}(n)$ considered above the group T can be thought of as the group of diagonal matrices; hence T is naturally isomorphic to $\mathbb{G}_m^n$. [1] Thus Bun_T classifies n-tuples $(\mathcal{L}_1, \ldots, \mathcal{L}_n)$ of line bundles on C. In terms of the above description of Bun_B the map q sends any flag $0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n$ to $(\mathcal{V}_1, \mathcal{V}_2/\mathcal{V}_1, \ldots, \mathcal{V}_n/\mathcal{V}_{n-1})$.

It is easy to see that in general the connected components of Bun_T are classified by elements of the lattice $\Lambda_G = \Lambda_T$. For each $\theta \in \Lambda_G$ we set $\mathrm{Bun}_B^{\theta} = q^{-1}(\mathrm{Bun}_T^{\theta})$. It is easy to see that the assignment $\theta \mapsto \mathrm{Bun}_B^{\theta}$ also defines a bijection between Λ_G and the set of connected components of Bun_B.

For each $\theta \in \Lambda_G$ the map $p\colon \mathrm{Bun}_B^{\theta} \to \mathrm{Bun}_G$ is representable. Moreover, it is clear that the fiber of this map over the trivial bundle in Bun_G is exactly our space $\mathrm{Maps}^{\theta}(C, X_{G,B})$ (note that the stack Bun_B^{θ} exists for any $\theta \in \Lambda_G$ but its fiber over the trivial bundle is non-empty only if $\theta \in \Lambda_G^+$). In general, the fibers of p (for fixed θ) are quasi-projective (but not projective) varieties; for various purposes (discussed, in particular, in other parts of this paper) it is useful to have a relative compactification $\overline{\mathrm{Bun}}_B^{\theta} \to \mathrm{Bun}_G$ such that its fiber over the trivial bundle in Bun_G will be exactly $\mathrm{QMaps}^{\theta}(C, X_{G,B})$. Such a compactification indeed can be constructed; let us give its explicit description (in particular, this will give a slightly different (but equivalent) definition of $\mathrm{QMaps}^{\theta}(C, X_{G,B})$.

We want to define $\overline{\mathrm{Bun}}_B$ as a solution to some moduli problem. Since $\overline{\mathrm{Bun}}_B$ is going to be an algebraic stack we must define the *groupoid* of S-points of $\overline{\mathrm{Bun}}_B$ for any scheme S over $\mathbb{C}$.

Let $\check{\Lambda}_G$ be the dual lattice of Λ_G. This is the weight lattice of the group G [2] We

[1] Here $\mathbb{G}_m$ denotes the multiplicative group

[2] The reader may find this notation a bit bizarre, since usually one uses the $\check{}$-notation for coweights and not for weights. However, it turns out that here it is much more convenient to use our notation; the main reason for this

define an S-point of $\overline{\mathrm{Bun}}_B$ to be a triple $(\mathcal{F}_G, \mathcal{F}_T, \kappa^{\check{\lambda}},\ \forall \check{\lambda} \in \check{\Lambda}_G^+)$, where $\mathcal{F}_G$ and $\mathcal{F}_T$ are as above, and $\kappa^{\check{\lambda}}$ is a map of coherent sheaves

$$\mathcal{L}^{\check{\lambda}}_{\mathcal{F}_T} \hookrightarrow \mathcal{V}^{\check{\lambda}}_{\mathcal{F}_G},$$

such that for every geometric point $s \in S$ the restriction $\kappa^{\check{\lambda}}|_{X\times s}$ is an injection. The last condition is equivalent to saying that $\kappa^{\check{\lambda}}$ is an injection such that the quotient $\mathcal{V}^{\check{\lambda}}_{\mathcal{F}_G}/\operatorname{Im}(\kappa^{\check{\lambda}})$ is S-flat.

The system of embeddings $\kappa^{\check{\lambda}}$ must satisfy the so-called *Plücker relations* which can be formulated as follows.

First, for $\check{\lambda} = 0$, κ^0 must be the identity map $\mathcal{O} \simeq \mathcal{L}^0_{\mathcal{F}_T} \to \mathcal{V}^0_{\mathcal{F}_G} \simeq \mathcal{O}$. Secondly, for two dominant integral weights $\check{\lambda}$ and $\check{\mu}$, the map

$$\mathcal{L}^{\check{\lambda}}_{\mathcal{F}_T} \otimes \mathcal{L}^{\check{\mu}}_{\mathcal{F}_T} \xrightarrow{\kappa^{\check{\lambda}}\otimes\kappa^{\check{\mu}}} \mathcal{V}^{\check{\lambda}}_{\mathcal{F}_G} \otimes \mathcal{V}^{\check{\mu}}_{\mathcal{F}_G} \simeq (\mathcal{V}^{\check{\lambda}} \otimes \mathcal{V}^{\check{\mu}})_{\mathcal{F}_G}$$

must coincide with the composition

$$\mathcal{L}^{\check{\lambda}}_{\mathcal{F}_T} \otimes \mathcal{L}^{\check{\mu}}_{\mathcal{F}_T} \simeq \mathcal{L}^{\check{\lambda}+\check{\mu}}_{\mathcal{F}_T} \xrightarrow{\kappa^{\check{\lambda}+\check{\mu}}} \mathcal{V}^{\check{\lambda}+\check{\mu}}_{\mathcal{F}_G} \to (\mathcal{V}^{\check{\lambda}} \otimes \mathcal{V}^{\check{\mu}})_{\mathcal{F}_G}.$$

It is easy to see that if all the maps $\kappa^{\check{\lambda}}$ are *embeddings of subbundles* (i.e. $\kappa^{\check{\lambda}}$ does not vanish on any fiber over any $c \in C$ then the collection $(\mathcal{F}_G, \mathcal{F}_T)$ together with all $\kappa^{\check{\lambda}}$ defines a point of Bun_B.

Here is another (somewhat more geometric) definition of $\overline{\mathrm{Bun}}_B$ (note that restricting to the fiber over the trivial bundle we get yet another definition of $\mathrm{QMaps}^{\theta}(C, X)$.

Let us denote by $U \subset B$ the unipotent radical of B. Since we have the natural isomorphism $G/U = T$ it follows that the variety G/U is endowed with a natural right action of T (of course, it also has a natural left G-action).

It is now easy to see that the stack Bun_B classifies the following data:

$$(\mathcal{F}_G;\, \mathcal{F}_T;\, \kappa\colon \mathcal{F}_G \to G/U \overset{T}{\times} \mathcal{F}_T),$$

where $\mathcal{F}_G$ is a G-bundle, $\mathcal{F}_T$ is a T-bundle and κ is a G-equivariant map.

Recall that G/U is a quasi-affine variety and let $\overline{G/U}$ denote its affine closure. The groups G and T act on G/U and therefore also on $\overline{G/U}$. The basic example of these varieties that one should keep in mind is the case $G = \mathrm{SL}(2)$. In this case G/U can be naturally identified with $\mathbb{A}^2\backslash\{0\}$ and $\overline{G/U} = \mathbb{A}^2$ (here $\mathbb{A}^2$ denotes the affine plane).

We claim now that an S-point of $\overline{\mathrm{Bun}}_B$ is the same as a triple $(\mathcal{F}_G, \mathcal{F}_T, \kappa)$, where $\mathcal{F}_G$ (resp., $\mathcal{F}_T$) is an S-point of Bun_G (resp., of Bun_T) and κ is a G-equivariant map

$$\mathcal{F}_G \to \overline{G/U} \overset{T}{\times} \mathcal{F}_T,$$

comes from the fact that many results will be formulated in terms of the Langlands dual group $\check{G}$ whose *weight* lattice is Λ_G!

such that for every geometric point $s \in S$ there is a Zariski-open subset $C^0 \subset C \times s$ such that the map

$$\kappa|_{C^0} : \mathcal{F}_G|_{C^0} \to \overline{G/U} \overset{T}{\times} \mathcal{F}_T|_{C^0}$$

factors through $G/U \overset{T}{\times} \mathcal{F}_T|_{C^0} \subset \overline{G/U} \overset{T}{\times} \mathcal{F}_T|_{C^0}$.

2.5. Quasi-maps into partial flag varieties. Let now $P \subset G$ be an arbitrary parabolic subgroup of G. Then as before we may consider the stack Bun_P of principal P-bundles on C; this stack is again naturally mapped to Bun_G and we would like to find some natural relative compactification of it. It turns out that in this case there exist *two different* natural compactifications $\overline{\mathrm{Bun}}_P$ and $\widetilde{\mathrm{Bun}}_P$ such that the embedding of Bun_P into both of them extends to a projective morphism $\widetilde{\mathrm{Bun}}_P \to \overline{\mathrm{Bun}}_P$. We refer the reader to [10] for the corresponding definitions. Here we shall only explain the geometric source for the existence of two such compactifications.

As was explained above the stacks $\overline{\mathrm{Bun}}_B$ are closely related with the varieties G/U and their affine closures $\overline{G/U}$; it is of crucial importance that G/U has a free T-action such that $(G/U)/T = G/B$.

Given a parabolic subgroup P as above one can attach two quasi-affine G-varieties to it: the first one is $G/[P, P]$ and the second one is G/U_P (here U_P denotes the unipotent radical of P; note that if P is a Borel subgroup of G then $[P, P] = U_P$). Let M denote the Levi group of P; by definition $M = P/U_P$. Also, one has the natural isomorphism $P/[P, P] = M/[M, M]$. Thus the first variety has a natural free action of $M/[M, M]$ and the second has an action of M; moreover, one has $(G/[P, P])/(M/[M, M]) = G/P = (G/U_P)/M$. Thus one can use the quasi-affine closures of $G/[P, P]$ and of G/U_P to construct two relative compactifications $\overline{\mathrm{Bun}}_P$ and $\widetilde{\mathrm{Bun}}_P$ of the stack Bun_P in the way similar to what was explained above for $P = B$.

Taking the fibers of the above stacks over the trivial bundle in Bun_G we get two different versions of quasi-maps from C to $G/P = X_{G,P}$. In what follows we shall denote by $\mathcal{QM}^{\theta}_{G,P}$ the space of quasi-maps $\mathbb{P}^1 \to X_{G,P}$ coming from $\overline{\mathrm{Bun}}_P$ (the compactification having to do with the variety G/P). Here θ should be a positive element of the lattice $\Lambda_{G,M}$ which is the lattice of cocharacters of $M/[M, M]$.

It turns out that many of the above definitions may be given also when $\mathfrak{g}$ is replaced by an affine Kac–Moody Lie algebra; the corresponding spaces of maps and quasi-maps are closely related to moduli spaces of G-bundles on a rational algebraic surface. This will be discussed in Section 5 (for more details the reader should consult [8]).

3. Quasi-maps into flag varieties and semi-infinite Schubert varieties

3.1. Ordinary Schubert varieties and their singularities. Let G as a before be a semi-simple simply connected algebraic group and let B be a Borel subgroup of it.

Recall that we denote $X_{G,B} = G/B$. It is well known that the set of B-orbits on $X_{G,B}$ is in one-to-one correspondence with the elements of the Weyl group W of G. For each $w \in W$ we denote the corresponding orbit by $X^w_{G,B}$. It is also known that each $X^w_{G,B}$ is isomorphic to the affine space $\mathbb{A}^{\ell(w)}$ where $\ell\colon W \to \mathbb{Z}_+$ is the length function.

The closure $\overline{X}^w_{G,B} \subset X_{G,B}$ of $X^w_{G,B}$ is usually called the Schubert variety attached to W. The singularities of these varieties play a very important role in various branches of representation theory. It is known (cf. [14] and references therein) that these varieties are normal and have rational singularities. Let $\mathrm{IC}^w_{G,B}$ denote the intersection cohomology sheaf of $X^w_{G,B}$. It is also well known that the stalks of $\mathrm{IC}^w_{G,B}$ can be described in terms of the *Kazhdan–Lusztig polynomials* attached to W (cf. [29]).

More generally, given two parabolic subgroups $P, Q \subset G$ one may study the closures of Q-orbits on G/P; these are the most general *parabolic Schubert varieties*. The stalks of their IC-sheaves are computed by *parabolic Kazhdan–Lusztig polynomials*.

One can generalize the above construction to the loop (or affine) groups associated with G. Namely, given a parabolic subgroup P as above one may construct two different affine flag varieties $X^{\mathrm{aff}}_{G,P}$ and $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ associated with the pair G, P. We shall refer to the first one as the corresponding *thin flag affine partial flag variety* and to the second one as the *thick partial affine flag variety* (in principle, there exist more general partial affine flag varieties but we shall never consider them in this paper). Set-theoretically, we can describe $X^{\mathrm{aff}}_{G,P}$ and $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ as follows.

Let $\mathcal{K} = \mathbb{C}((t))$ be the field of formal Laurent power series; let $\mathcal{O} \subset \mathcal{K}$ be the ring of Taylor series. Consider the "loop" group $G(\mathcal{K})$ of $\mathcal{K}$-points of G. Let $I_P \subset G(\mathcal{K})$ denote the subgroup of $G(\mathcal{O}) \subset G(\mathcal{K})$ consisting of those Taylor series whose value at $t = 0$ lies in $P \subset G$. When $P = B$ is a Borel subgroup we shall write just I instead of I_B and call it the Iwahori subgroup of $G((t))$. We shall also denote by $I^0 \subset I$ its pro-unipotent radical (it consists of those Taylor series as above whose value at $t = 0$ lies in the unipotent radical U of B). Note also that when $P = G$ we have $I_G = G(\mathcal{O})$.

Similarly, we can define the group $\boldsymbol{I}_P \subset G[t^{-1}]$ consisting of those polynomials in t^{-1} whose value at $t = \infty$ lies in P. Thus on level of $\mathbb{C}$-points we have

$$X^{\mathrm{aff}}_{G,P} = G(\mathcal{K})/I_P; \quad \boldsymbol{X}^{\mathrm{aff}}_{G,P} = G(\mathcal{K})/\boldsymbol{I}_P.$$

We shall be mostly talking about Schubert varieties in $X^{\mathrm{aff}}_{G,P}$ (the flag varieties $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ will also appear in Section 5 of this paper). By definition, these are closures of I_Q-orbits in some $X^{\mathrm{aff}}_{G,P}$. These are known to be finite-dimensional normal projective varieties having rational singularities (cf. [14]). In the case when $P = Q = B$ (i.e. when we are dealing with I-orbits on $X^{\mathrm{aff}}_{G,B}$) the orbits are classified by elements of the affine Weyl group W_{aff} (by definition, this is the semi-direct product of the Weyl group W of G and the lattice Λ_G). At the other extreme, when $P = Q = G$ we are dealing with $G(\mathcal{O})$-orbits on $G(\mathcal{K})/G(\mathcal{O})$; these orbits are in one-to-one correspondence

with Λ_G/W. The latter set can be identified with the set of dominant weights of the Langlands dual group $\check{G}$.

One of the reasons that the complete flag variety $X_{G,B}$ plays a distinguished role in representation theory is the Beilinson–Bernstein localization theorem (cf. [4]) which allows one to realize representations of the Lie algebra $\mathfrak{g}$ (with fixed central character of the universal enveloping algebra $U(\mathfrak{g})$) in terms of algebraic D-modules on $X_{G,B}$. In this way the category of B-equivariant (or, more generally, U-equivariant) modules corresponds to the regular block of the so-called category $\mathcal{O}$. Similar (but much less understood) statements hold in the affine case too. Namely let $\mathfrak{g}_{\mathrm{aff}}$ denote the affine Lie algebra corresponding to $\mathfrak{g}$. By definition this algebra is a central extension of the loop algebra $\mathfrak{g}((t))$:

$$0 \to \mathbb{C} \to \mathfrak{g}_{\mathrm{aff}} \to \mathfrak{g}((t)) \to 0.$$

Then one gets a geometric realization (of some part of) the category for the corresponding affine Lie algebra $\mathfrak{g}_{\mathrm{aff}}$; the variety $X^{\mathrm{aff}}_{G,B}$ allows to realize $\mathfrak{g}_{\mathrm{aff}}$-modules at the negative level and the variety $\boldsymbol{X}^{\mathrm{aff}}_{G,B}$ has to do with $\mathfrak{g}_{\mathrm{aff}}$-modules on the positive level. We refer the reader to [27] and references therein for more details.

3.2. The semi-infinite flag manifold. The semi-infinite flag manifold $X^{\frac{\infty}{2}}_{G,B}$ is usually defined as the quotient $G(\mathcal{K})/T(\mathcal{O}) \cdot U(\mathcal{K})$. In terms of algebraic geometry this "space" seems to be widely infinite-dimensional. However, one still would like to think of it as some kind of geometric object; this should have many applications to representation theory.

More specifically, we would like to mention the following two problems:

1) Construct the category of D-modules (or perverse sheaves) on $X^{\frac{\infty}{2}}_{G,B}$ and relate it to some other abelian categories coming from representation theory of affine Lie algebras and quantum groups.

2) It is easy to see that the orbits of the Iwahori group I on $X^{\frac{\infty}{2}}_{G,B}$ are classified by elements of the affine Weyl group W_{aff} attached to G. We shall denote by $X^{w,\frac{\infty}{2}}_{G,B}$ the orbit corresponding to $w \in W_{\mathrm{aff}}$. Then the problem reads as follows: explain in what sense the singularities of the closures of $\overline{X}^{w,\frac{\infty}{2}}_{G,B}$ are finite-dimensional and "understand" those singularities. In particular, one should be able to compute the stalks of the IC-sheaves associated to those singularities and relate them to the periodic polynomials defined in [36]. More generally, one can study I_P-orbits on $X^{\frac{\infty}{2}}_{G,B}$ together with their closures. We shall refer to them as the (not yet constructed) *semi-infinite Schubert varieties*.

In particular, the $G(\mathcal{O})$-orbits on $X^{\frac{\infty}{2}}_{G,B}$ are classified by elements of Λ_G; for each $\mu \in \Lambda_G$ we shall denote the corresponding orbit simply by S^μ. It is easy to see that if the closure $\overline{S}^\mu$ of S^μ makes any reasonable sense then it must be equal to the union of all S^ν's with $\nu - \mu \in \Lambda^+_G$. Also, the lattice $\Lambda_G = T(\mathcal{K})/T(\mathcal{O})$ acts on $X^{\frac{\infty}{2}}_{G,B}$ on the right and every $\gamma \in \Lambda_G$ maps the orbit S^μ to $S^{\mu+\gamma}$. Hence the singularity of $\overline{S}^\mu$

in the neighborhood of a point of S^ν depends only on $\nu - \mu \in \Lambda_G^+$. In particular, if the intersection cohomology sheaf $\mathrm{IC}(\overline{S}^\mu)$ makes sense, then its stalk at a point of S^ν should only depend on $\nu - \mu = \theta$. In fact, from the results of [35], [36] and references therein it is natural to expect that this stalk comes from the graded vector space U_θ computed as follows: let $\check{\mathfrak{g}}$ denote the Langlands dual Lie algebra whose root system is dual to that of $\mathfrak{g}$. We have its triangular decomposition $\check{\mathfrak{g}} = \check{\mathfrak{n}}_- \oplus \check{\mathfrak{t}} \oplus \check{\mathfrak{n}}_+$. Also $\check{\mathfrak{t}} = \mathfrak{t}^*$ and we may identify Λ_G with the root lattice $\check{\mathfrak{g}}$. Consider the symmetric algebra $\mathrm{Sym}(\mathfrak{n}_+)$ with the natural even grading on it (defined by the requiring that the subspace $\check{\mathfrak{n}}_+ \subset \mathrm{Sym}(\mathfrak{n}_+)$ has degree 2). The dual Cartan torus $\check{T}$ acts on this algebra (since it acts on $\mathfrak{n}_+$); for each $\theta \in \Lambda_G^+$ we may consider the subspace $\mathrm{Sym}(\mathfrak{n}_+)_\theta \subset \mathrm{Sym}(\mathfrak{n}_+)$ on which $\check{T}$ acts by the character θ. This space inherits the grading from $\mathrm{Sym}(\mathfrak{n}_+)$. Let also $\check{\rho} \in \check{\Lambda}_G$ denote the half-sum of the positive roots of G. Then, guided by the results of *loc. cit.* one expects to have

$$U_\theta \simeq \mathrm{Sym}(\mathfrak{n}_+)_\theta[2\langle\theta, \check{\rho}\rangle]. \tag{3.1}$$

The general principle (due to Drinfeld) says that one should be able to use the quasi-maps spaces $\mathcal{QM}^\theta_{G,B}$ (or, the stacks $\overline{\mathrm{Bun}}_B$ for any smooth projective curve C) as "finite-dimensional models" for the semi-infinite flag manifolds and the semi-infinite Schubert varieties. In particular, one expects to be able construct the correct category of D-modules using quasi-maps as well as to turn (3.1) into a mathematical theorem. This has indeed been performed in the works [1], [15] and [9]. Let us give a brief sketch of the results of *loc. cit.*

3.3. Localization theorem for the small quantum group. Let us turn to some illustrations of the above principle. First of all, in [1] we propose a definition of the category $\mathrm{Perv}(X^{\frac{\infty}{2}}_{G,B})$ in terms of the stacks $\overline{\mathrm{Bun}}_B$. We give a representation-theoretic interpretation of the corresponding subcategory $\mathrm{Perv}_{I^0}(X^{\frac{\infty}{2}}_{G,B})$ consisting of I^0-equivariant perverse sheaves; it turns out to be equivalent to the *regular block* of category of graded representations of the so-called *small quantum group* $\mathfrak{u}_\ell$ attached to the Lie algebra $\mathfrak{g}$; here ℓ denotes a root of unity satisfying some mild assumptions (cf. [1] for more details). This result was conjectured by B. Feigin in the early 90s. Another representation-theoretic interpretation of the same category (in terms of representations of the affine Lie algebra $\mathfrak{g}_{\mathrm{aff}}$) should appear soon in the works of Frenkel and Gaitsgory.

3.4. Computation of the IC-sheaf. Another check of the above principle will be to compute the stalks of the IC-sheaves of the spaces $\mathcal{QM}^\theta_{G,B}$ (or the stacks $\overline{\mathrm{Bun}}_B$) and compare it with (3.1). This was done in [15]; also in [9] this was generalized to arbitrary parabolic $P \subset G$. More specifically, the space $\mathcal{QM}^\theta_{G,B}$ possesses the

following stratification (similar to (2.1)):

$$\mathcal{QM}^{\theta}_{G,B} = \bigcup_{\mu\in\Lambda^+_G} \mathcal{M}^{\theta-\mu}_{G,B} \times \mathrm{Sym}^{\mu}(\mathbb{P}^1). \tag{3.2}$$

Here by $\mathrm{Sym}^{\mu}(\mathbb{P}^1)$ we mean the space of all colored divisors $\sum \mu_i x_i$ where $\mu_i \in \Lambda^+_G$, $x_i \in \mathbb{P}^1$ and $\sum \mu_i = \mu$. Then we have

Theorem 3.1. *The stalk of* $\mathrm{IC}_{\mathcal{QM}^{\theta}_{G,B}}$ *at a point of* $\mathcal{QM}^{\theta}_{G,B}$ *corresponding to a colored divisor* $\sum \mu_i x_i$ *as above is equal to* $\otimes_i \mathrm{Sym}(\check{\mathfrak{n}}_+)_{\mu_i}[2\langle\mu_i, \check{\rho}\rangle]$. *In particular, the stalk of* $\mathrm{IC}_{\mathcal{QM}^{\theta}_{G,B}}$ *at the "most singular" points of* $\mathcal{QM}^{\theta}_{G,B}$ *corresponding to the divisor of the form* $\theta \cdot x$ *(for some* $x \in \mathbb{P}^1$*) is equal to* $U_\theta = \mathrm{Sym}(\mathfrak{n}_+)_\theta[2\langle\theta, \check{\rho}\rangle]$.

The proof of Theorem 3.1 relies on many things, in particular the results of [37] about semi-infinite orbits in the affine Grassmannian of G.

3.5. Geometric construction of the universal Verma module. We have seen that one can read off some information related to the Langlands dual Lie algebra from the singularities of the quasi-maps' spaces. It is natural to ask if one could push this a little further and get a geometric construction of $\check{\mathfrak{g}}$-modules (in Section 5 we are going to generalize it to affine Lie algebras).

Of course, the most interesting modules that one would like to get in this way are the finite-dimensional modules. This, however, has not been done yet. In this section we explain how to use the spaces of quasi-maps in order to construct the "universal Verma module" for the Lie algebra $\check{\mathfrak{g}}$. We also give geometric interpretation of the Shapovalov form and the Whittaker vectors (cf. the definitions below). We shall generalize this in Subsection 5.3 to the case of affine Lie algebras. These constructions will play the crucial role in Section 6 where we discuss applications of our techniques to some questions of enumerative algebraic geometry.

First, let Y be a scheme endowed with an action of a reductive algebraic group L (in most applications L will actually be a torus). We denote by $\mathrm{IH}_L(Y)$ the intersection cohomology of Y with complex coefficients. This is a module over the algebra $\mathcal{A}_L = H^*_L(pt)$ which is known to be isomorphic to the algebra of polynomial functions on the Lie algebra $\mathfrak{l}$ of L which are invariant under the adjoint action of L. We let $\mathcal{K}_L$ denote the field of fractions of $\mathcal{A}_L$.

We now take Y to be the space ${}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,B}$ of *based* quasi-maps $\mathbb{P}^1 \to X_{G,B}$. By definition, this is the locally closed subscheme of $\mathcal{QM}^{\theta}_{G,B}$ corresponding to those quasi-maps which are first of all well-defined as *maps* around $\infty \in \mathbb{P}^1$ and such that their value at ∞ is equal to the point $e_{G,B} \in X_{G,B}$ corresponding to the unit element of G under the identification $X_{G,B} = G/B$. This scheme is endowed with a natural action of the torus $T \times \mathbb{C}^*$ (here T acts on $X_{G,B}$ preserving $e_{G,B}$ and $\mathbb{C}^*$ acts on $\mathbb{P}^1$

preserving ∞). Define

$$\mathrm{IH}^{\theta}_{G,B} = \mathrm{IH}^{*}_{T\times\mathbb{C}^*}({}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,B}) \underset{\mathcal{A}_{T\times\mathbb{C}^*}}{\otimes} \mathcal{K}_{T\times\mathbb{C}^*}, \quad \mathrm{IH}_{G,B} = \bigoplus_{\theta\in\Lambda^{+}_{G,B}} \mathrm{IH}^{\theta}_{G,B}.$$

Each $\mathrm{IH}^{\theta}_{G,B}$ is a finite-dimensional vector space over the field $\mathcal{K}_{T\times\mathbb{C}^*}$ which can be thought of as the field of rational functions of the variables $a \in \mathfrak{t}$ and $\hbar \in \mathbb{C}$. Moreover, $\mathrm{IH}^{\theta}_{G,B}$ is endowed with a (non-degenerate) Poincaré pairing $\langle\cdot,\cdot\rangle^{\theta}_{G,P}$ taking values in $\mathcal{K}_{T\times\mathbb{C}^*}$ (one has to explain why the Poincaré pairing is well defined since ${}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,B}$ is not projective; this is a corollary of (some version of) the localization theorem in equivariant cohomology – cf. [5] for more details).

In [5] we construct a natural action of the Lie algebra $\check{\mathfrak{g}}$ on the space $\mathrm{IH}_{G,B}$. Moreover, this action has the following properties. First of all, let us denote by $\langle\cdot,\cdot\rangle_{G,B}$ the direct sum of the pairings $(-1)^{\langle\theta,\check{\rho}\rangle}\langle\cdot,\cdot\rangle^{\theta}_{G,B}$.

Recall that the Lie algebra $\check{\mathfrak{g}}$ has its triangular decomposition $\check{\mathfrak{g}} = \check{\mathfrak{n}}_+ \oplus \check{\mathfrak{t}} \oplus \check{\mathfrak{n}}_-$. Let $\kappa\colon \check{\mathfrak{g}} \to \check{\mathfrak{g}}$ denote the Cartan anti-involution which interchanges $\check{\mathfrak{n}}_+$ and $\check{\mathfrak{n}}_-$ and acts as identity on $\check{\mathfrak{t}}$. For each $\lambda \in \mathfrak{t} = (\check{\mathfrak{t}})^*$ we denote by $M(\lambda)$ the corresponding Verma module with lowest weight λ; this is a module generated by a vector v_λ with (the only) relations

$$t(v_\lambda) = \lambda(t)v_\lambda \ \text{ for } t \in \check{\mathfrak{t}} \quad \text{and} \quad n(v_\lambda) = 0 \ \text{ for } n \in \check{\mathfrak{n}}_-.$$

Then:

1) $\mathrm{IH}_{G,B}$ (with the above action) becomes isomorphic to $M(\lambda)$ where $\lambda = \frac{a}{\hbar} + \rho$.

2) $\mathrm{IH}^{\theta}_{G,B} \subset \mathrm{IH}_{G,B}$ is the $\frac{a}{\hbar} + \rho + \theta$-weight space of $\mathrm{IH}_{G,B}$.

3) For each $g \in \check{\mathfrak{g}}$ and $v, w \in \mathrm{IH}_{G,B}$ we have

$$\langle g(v), w\rangle_{G,B} = \langle v, \kappa(g)w\rangle_{G,B}.$$

4) The vector $\sum_\theta 1^{\theta}_{G,B}$ (lying in the corresponding completion of $\mathrm{IH}_{G,B}$) is a Whittaker vector (i.e. an $\mathfrak{n}_-$-eigen-vector) for the above action.

We are not going to explain the construction of the action in this survey paper. Let us only make a few remarks about it. In the case $G = \mathrm{SL}(n)$ the smallness result of [30] allows to replace the intersection cohomology of ${}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,B}$ by the ordinary cohomology of the corresponding based version of the Laumon resolution ${}^{\mathrm{b}}\mathcal{QM}^{L,\theta}_{G,B}$; on the latter (equivariant localized) cohomology the action of the Chevalley generators of $\check{\mathfrak{g}} = \mathfrak{sl}(n)$ can be defined by means of some explicit correspondences (this is similar to the main construction of [17]; also in [7] we generalize this to the case when equivariant cohomology is replaced by equivariant K-theory. In this case the action of the Lie algebra $\mathfrak{sl}(n)$ is replaced by the action of the corresponding quantum group $U_q(\mathfrak{sl}(n))$). Also for any G the fact, that the dimension of $\mathrm{IH}^{\theta}_{G,B}$ can be easily deduced from Theorem 3.1. Our construction of the $\check{\mathfrak{g}}$-action on $\mathrm{IH}_{G,B}$ is very close to the construction in Section 4 of [15].

4. The stack $\overline{\mathrm{Bun}}_B$ and geometric Eisenstein series

This section is devoted to an application of the stacks $\overline{\mathrm{Bun}}_B$ to some questions of *geometric Langlands correspondence*. A reader who is not interested in the subject may skip this section since it will never be used in the future. In fact we are going to discuss only one such application (which was the first one historically) – the construction of *geometric Eisenstein series*. Let us note, though, that the stacks $\overline{\mathrm{Bun}}_B$ have appeared in many other works on the subject. For example they play the crucial role in the geometric proof of Casselman–Shalika formula by E. Frenkel, D. Gaitsgory and K. Vilonen (cf. [18]) as well as in D. Gaitsgory's work (cf. [20]) on the so-called "vanishing conjecture" which implies (the main portion of) the geometric Langlands conjecture for $\mathrm{GL}(n)$, as well as de Jong's conjecture about representations of Galois groups of functional fields (cf. [22]). A good review of these results may be found in [21].

All the results discussed below are taken from [10].

4.1. The usual Eisenstein series. Let X be a curve over $\mathbb{F}_q$ and let G be a reductive group. The classical theory of automorphic forms is concerned with the space of functions on the quotient $G_{\mathbb{A}}/G_{\mathcal{K}}$, where $\mathcal{K}$ (resp., $\mathbb{A}$) is the field or rational functions on X (resp., the ring of adèles of $\mathcal{K}$). In this paper, we will consider only the unramified situation, i.e. we will study functions (and afterwards perverse sheaves) on the double quotient $G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}$.

Let T be a Cartan subgroup of G. There is a well-known construction, called the *Eisenstein series operator* that attaches to a compactly supported function on $T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}}$ a function on $G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}$:

Consider the diagram

$$\begin{array}{ccc} B_{\mathbb{O}}\backslash B_{\mathbb{A}}/B_{\mathcal{K}} & \xrightarrow{q} & T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}} \\ \downarrow{\scriptstyle p} & & \\ G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}, & & \end{array}$$

where B is a Borel subgroup of G. Up to a normalization factor, the Eisenstein series of a function S on $T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}}$ is $p_!(q^*(S))$, where q^* denotes pull-back and $p_!$ is integration along the fiber.

Our goal is to study a geometric analog of this construction.

Let Bun_G denote the stack of G–bundles on X. One may regard the derived category of constructible sheaves on Bun_G (denoted $\mathrm{Sh}(\mathrm{Bun}_G)$) as a geometric analog of the space of functions on $G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}$. Then, by geometrizing the Eisenstein series operator, we obtain an Eisenstein series functor Eis' similar to the above one, where the intermediate stack is Bun_B – the stack of B – bundles on X.

However, this construction has an immediate drawback – it is not sufficiently functorial (for example it does not commute with Verdier duality), the reason being

that the projection $p\colon \mathrm{Bun}_B \to \mathrm{Bun}_G$ has non-compact fibers. Therefore, it is natural to look for a relative compactification of Bun_B along the fibers of the projection p.

It turns out that the compactification $\overline{\mathrm{Bun}}_B$ discussed indeed does the job, i.e. we can use it to define the corrected functor $\mathrm{Eis}\colon \mathrm{Sh}(\mathrm{Bun}_T) \to \mathrm{Sh}(\mathrm{Bun}_G)$. The paper [10] is devoted to the investigation of various properties of this functor.

In fact, all the technical results about the functor Eis essentially reduce to questions about the geometry of $\overline{\mathrm{Bun}}_B$ and the behaviour of the intersection cohomology sheaf on it.

We should say right away that the pioneering work in this direction was done by G. Laumon in [32], who considered the case of $G = \mathrm{GL}(n)$ using his own compactification $\overline{\mathrm{Bun}}_B^L$ of the stack Bun_B. In the sequel we will explain how the two approaches are related.

4.2. Survey of the main results of [10]. Once the stack $\overline{\mathrm{Bun}}_B$ is constructed, one can try to use it to define the "compactified" Eisenstein series functor $\mathrm{Eis}\colon \mathrm{Sh}(\mathrm{Bun}_T) \to \mathrm{Sh}(\mathrm{Bun}_G)$. Let p and q denote the natural projections from $\overline{\mathrm{Bun}}_B$ to Bun_G and Bun_T, respectively. The first idea would be to consider the functor $\mathcal{S} \in \mathrm{Sh}(\mathrm{Bun}_T) \mapsto p_!(q^*(\mathcal{S})) \in \mathrm{Bun}_G$. However, this is too naive, since if we want our functor to commute with Verdier duality, we need to take into account the singularities of $\overline{\mathrm{Bun}}_B$. Therefore, one introduces a *kernel* on $\overline{\mathrm{Bun}}_B$ given by its intersection cohomology sheaf. I.e., we define the functor Eis by

$$\mathcal{S} \mapsto p_!(q^*(\mathcal{S}) \otimes \mathrm{IC}_{\overline{\mathrm{Bun}}_B}),$$

up to a cohomological shift and Tate's twist. Similarly, one defines the functor $\mathrm{Eis}_M^G\colon \mathrm{Sh}(\mathrm{Bun}_M) \to \mathrm{Sh}(\mathrm{Bun}_G)$, where M is the Levi quotient of a parabolic P.

The first test whether our definition of the functor Eis is "the right one" would be the assertion that Eis (or more generally Eis_M^G) indeed commutes with Verdier duality. It can be shown that our Eis indeed passes this test.

Let us again add a comment of how the functor Eis is connected to Laumon's work. One can define functors $\mathrm{Eis}^L\colon \mathrm{Sh}(\mathrm{Bun}_T) \to \mathrm{Sh}(\mathrm{Bun}_G)$ using Laumon's compactification. (In the original work [32], Laumon did not consider Eis^L as a functor, but rather applied it to specific sheaves on Bun_T.) However, from the smallness result of [30] it follows that the functors Eis^L and Eis are canonically isomorphic.

Once we defined the functors $\mathrm{Eis} = \mathrm{Eis}_T^G\colon \mathrm{Sh}(\mathrm{Bun}_T) \to \mathrm{Sh}(\mathrm{Bun}_G)$, $\mathrm{Eis}_M^G\colon \mathrm{Sh}(\mathrm{Bun}_M) \to \mathrm{Sh}(\mathrm{Bun}_G)$ and a similar functor for M, $\mathrm{Eis}_T^M\colon \mathrm{Sh}(\mathrm{Bun}_T) \to \mathrm{Sh}(\mathrm{Bun}_M)$, it is by all means natural to expect that these functors compose nicely, i.e. that $\mathrm{Eis}_T^G \simeq \mathrm{Eis}_M^G \circ \mathrm{Eis}_T^M$.

For example, if instead of Eis we used the naive (uncompactified) functor Eis$'$, the analogous assertion would be a triviality, since $\mathrm{Bun}_B \simeq \mathrm{Bun}_P \underset{\mathrm{Bun}_M}{\times} \mathrm{Bun}_{B(M)}$, where $B(M)$ is the Borel subgroup of M.

The problem with our definition of Eis_M^G is that there is no map between the relevant compactifications, i.e. from $\overline{\mathrm{Bun}}_B$ to $\widetilde{\mathrm{Bun}}_P$. Nevertheless, the assertion that $\mathrm{Eis}_T^G \simeq \mathrm{Eis}_M^G \circ \mathrm{Eis}_T^M$ does hold. This in fact is a non-trivial theorem proved in [10].

Here are the main properties of the Eisenstein series functor.

4.3. Behaviour with respect to the Hecke functors. Classically, on the space of functions on the double quotient $G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}$ we have the action of $\otimes_{x\in X}\mathcal{H}_x(G)$, where x runs over the set of places of $\mathcal{K}$, and for each $x \in X$, $\mathcal{H}_x(G)$ denotes the corresponding spherical Hecke algebra of the group G.

Similarly, $\otimes_{x\in X}\mathcal{H}_x(T)$ acts on the space of functions on $T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}}$. In addition, for every x as above, there is a canonical homomorphism $\mathcal{H}_x(G) \to \mathcal{H}_x(T)$ described as follows:

Recall that there is a canonical isomorphism (due to Satake) between $\mathcal{H}_x(G)$ and the Grothendieck ring of the category of finite-dimensional representations of the Langlands dual group $\check{G}$. We have the natural restriction functor $\mathrm{Rep}(\check{G}) \to \mathrm{Rep}(\check{T})$, and our homomorphism $\mathcal{H}_x(G) \to \mathcal{H}_x(T)$ corresponds to the induced homomorphism $K(\mathrm{Rep}(\check{G})) \to K(\mathrm{Rep}(\check{T}))$ between Grothendieck rings.

The basic property of the Eisenstein series operators is that it intertwines the $\mathcal{H}_x(G)$-action on $G_{\mathbb{O}}\backslash G_{\mathbb{A}}/G_{\mathcal{K}}$ and the $\mathcal{H}_x(T)$-action on $T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}}$ via the above homomorphism.

Our result below is a reflection of this phenomenon in the geometric setting. Now, instead of the Hecke algebras, we have the action of the Hecke functors on $\mathrm{Sh}(\mathrm{Bun}_G)$. Namely, for $x \in X$ and an object $V \in \mathrm{Rep}(\check{G})$, one defines the *Hecke functor*

$$\mathcal{S} \mapsto {}_xH_G(V, \mathcal{S})$$

from $\mathrm{Sh}(\mathrm{Bun}_G)$ to itself. The existence of such functors comes from the so-called *geometric Satake isomorphism* – cf. [37] and references therein.

We claim that for any $\mathcal{S} \in \mathrm{Sh}(\mathrm{Bun}_T)$ we have

$${}_xH_G(V, \mathrm{Eis}(\mathcal{S})) \simeq \mathrm{Eis}({}_xH_T(\mathrm{Res}^G_T(V), \mathcal{S})).$$

This result is more or less equivalent to one of the main results of [37]. A similar statement holds for the non-principal Eisenstein series functor Eis^G_M.

As a corollary, we obtain that if $E_{\check{M}}$ is an $\check{M}$-local system on X and $\mathrm{Aut}_{E_{\check{M}}}$ is a perverse sheaf (or a complex of sheaves) on Bun_M, corresponding to it in the sense of the geometric Langlands correspondence, then the complex $\mathrm{Eis}^G_M(\mathrm{Aut}_{E_{\check{M}}})$ on Bun_G is a Hecke eigen-sheaf with respect to the induced $\check{G}$-local system.

In particular, we construct Hecke eigen-sheaves for those homomorphisms $\pi_1(X) \to \check{G}$, whose image is contained in a maximal torus of $\check{G}$.

4.4. The functional equation. It is well known that the classical Eisenstein series satisfy the functional equation. Namely, let χ be a character of the group $T_{\mathbb{O}}\backslash T_{\mathbb{A}}/T_{\mathcal{K}}$ and let $\mathrm{w} \in W$ be an element of the Weyl group. We can translate χ by means of w and obtain a new (Grössen)-character χ^{w}.

The functional equation is the assertion that the Eisenstein series corresponding to χ and χ^{w} are equal, up to a ratio of the corresponding L-functions.

Now let $\mathcal{S}$ be an arbitrary complex of sheaves on Bun_T and let $\mathrm{w}\cdot\mathcal{S}$ be its w-translate. One may wonder whether there is any relation between $\mathrm{Eis}(\mathcal{S})$ and $\mathrm{Eis}(\mathrm{w}\cdot\mathcal{S})$.

We single out a subcategory in $\mathrm{Sh}(\mathrm{Bun}_T)$, corresponding to sheaves which we call "regular", for which we answer the above question. We show that for a regular sheaf $\mathcal{S}$ we have

$$\mathrm{Eis}(\mathcal{S}) \simeq \mathrm{Eis}(\mathrm{w}\cdot\mathcal{S}).$$

(It is easy to see that one should not expect the functional equation to hold for non-regular sheaves.)

A remarkable feature of this assertion is that the L-factors that enter the classical functional equation have disappeared. An explanation of this fact is provided by the corresponding result from [10] which says that the definition of Eis via the intersection cohomology sheaf on $\overline{\mathrm{Bun}}_B$ already incorporates the L-function.

We remark that an assertion similar to the above functional equation should hold also for non-principal Eisenstein series. Unfortunately, this seems to be beyond the access of our methods.

Using the above results we obtain a proof of the following very special case of the Langlands conjecture. Namely, we prove that if we start with an unramified irreducible representation of $\pi_1(X)$ into $\check{G}$, such that $\pi_1(X)^{\mathrm{geom}}$ [3] maps to $\check{T} \subset \check{G}$, then there exists an unramified automorphic form on $G_{\mathbb{A}}$ which corresponds to this representation in the sense of Langlands.

This may be considered as an application of the machinery developed in [10] to the classical theory of automorphic forms.

5. Quasi-maps into affine flag varieties and Uhlenbeck compactifications

In this section we take the base field to be $\mathbb{C}$.

5.1. The problem. Let G be an almost simple simply connected group over $\mathbb{C}$, with Lie algebra $\mathfrak{g}$, and let $\boldsymbol{S}$ be a smooth projective surface.

Let us denote by $\mathrm{Bun}_G^d(\boldsymbol{S})$ the moduli space (stack) of principal G-bundles on $\boldsymbol{S}$ of second Chern class $d \in \mathbb{Z}$. It is easy to see that $\mathrm{Bun}_G^d(\boldsymbol{S})$ cannot be compact and for many reasons it is natural to expect that there exists a compactification of $\mathrm{Bun}_G^d(\boldsymbol{S})$ which looks like a union

$$\bigcup_{b\in\mathbb{N}} \mathrm{Bun}_G^{d-b}(\boldsymbol{S}) \times \mathrm{Sym}^b(\boldsymbol{S}). \tag{5.1}$$

Note the striking similarity between (5.1) and (2.1).

[3] Here $\pi_1(X)^{\mathrm{geom}}$ denotes the geometric fundamental group of X – i.e. the fundamental group of X over the algebraic closure of $\mathbb{F}_q$

In the differential-geometric framework of moduli spaces of K-instantons on Riemannian 4-manifolds (where K is the maximal compact subgroup of G) such a compactification was introduced in the pioneering work [45]. Therefore, we shall call its algebro-geometric version the Uhlenbeck space, and denote it by $\mathcal{U}_G^d(\boldsymbol{S})$.

Unfortunately, one still does not know how to construct the spaces $\mathcal{U}_G^d(\boldsymbol{S})$ for a general group G and an arbitrary surface $\boldsymbol{S}$. More precisely, one would like to formulate a moduli problem, to which $\mathcal{U}_G^d(\boldsymbol{S})$ would be the answer, and so far this is not known. In this formulation the question of constructing the Uhlenbeck spaces has been posed (to the best of our knowledge) by V. Ginzburg. He and V. Baranovsky (cf. [3]) have made the first attempts to solve it, as well as indicated the approach adopted in this paper. The reader may also consult [2] for a different algebro-geometric approach to Uhlenbeck spaces.

A significant simplification occurs for $G = \mathrm{SL}_n$. Let us note that when $G = \mathrm{SL}_n$, there exists another natural compactification of the stack $\mathrm{Bun}_n^d(\boldsymbol{S}) := \mathrm{Bun}_{\mathrm{SL}_n}^d(\boldsymbol{S})$ (called the Gieseker compactification), by torsion-free sheaves of generic rank n and of second Chern class a, called the Gieseker compactification, which in this paper we will denote by $\widetilde{\mathfrak{N}}_n^d(\boldsymbol{S})$. One expects that there exists a proper map $\mathsf{f}\colon \widetilde{\mathfrak{N}}_n^d(\boldsymbol{S}) \to \mathcal{U}_{\mathrm{SL}_n}^d(\boldsymbol{S})$, described as follows:

A torsion-free sheaf $\mathcal{M}$ embeds into a short exact sequence

$$0 \to \mathcal{M} \to \mathcal{M}' \to \mathcal{M}_0 \to 0,$$

where $\mathcal{M}'$ is a vector bundle (called the saturation of $\mathcal{M}$), and $\mathcal{M}_0$ is a finite-length sheaf. The map should send a point of $\widetilde{\mathfrak{N}}_n^d(\boldsymbol{S})$ corresponding to $\mathcal{M}$ to the pair $(\mathcal{M}', \mathrm{cycle}(\mathcal{M}_0)) \in \mathrm{Bun}_n^{d-b}(\boldsymbol{S}) \times \mathrm{Sym}^b(\boldsymbol{S})$, where b is the length of $\mathcal{M}_0$, and $\mathrm{cycle}(\mathcal{M}_0)$ is the cycle of $\mathcal{M}_0$. In other words, the map must "collapse" the information of the quotient $\mathcal{M}' \to \mathcal{M}_0$ to just the information of the length of $\mathcal{M}_0$ at various points of $\boldsymbol{S}$.

Since the spaces $\widetilde{\mathfrak{N}}_n^d(\boldsymbol{S})$, being a solution of a moduli problem, are easy to construct, one may attempt to construct the Uhlenbeck spaces $\mathcal{U}_{\mathrm{SL}_n}^d(\boldsymbol{S})$ by constructing an explicit blow down of the Gieseker spaces $\widetilde{\mathfrak{N}}_n^d(\boldsymbol{S})$. This has indeed been performed in the works of J. Li (cf. [33]) and J. W. Morgan (cf. [38]).

The problem simplifies even further, when we put $\boldsymbol{S} = \mathbb{P}^2$, the projective plane, and consider bundles trivialized along a fixed line $\mathbb{P}^1 \subset \mathbb{P}^2$. In this case, the sought-for space $\mathcal{U}_n^d(\boldsymbol{S})$ has been constructed by S. Donaldson and thoroughly studied by H. Nakajima (cf. e.g. [39]) in his works on quiver varieties.

In [8] we consider the case of an arbitrary group G, but the surface equal to $\mathbb{P}^2$ (and we will be interested in bundles trivialized along $\mathbb{P}^1 \subset \mathbb{P}^2$, i.e., we will work in the Donaldson–Nakajima set-up.)

In fact we are able to construct the Uhlenbeck spaces $\mathcal{U}_G^d$, but only up to nilpotents. I.e., we will have several definitions, two of which admit modular descriptions, and which produce the same answer on the level of reduced schemes. We do not know, whether the resulting schemes actually coincide when we take the nilpotents into

account. And neither do we know whether the resulting reduced scheme is normal.

We should say that the problem of constructing the Uhlenbeck spaces can be posed over a base field of any characteristic. However, the proof of one the main results of [10], which insures that our spaces $\mathcal{U}_G^d$ are invariantly defined, uses the char $= 0$ assumption. It is quite possible that in order to treat the char $= p$ case, one needs a finer analysis.

5.2. A sketch of the construction. The construction of $\mathcal{U}_G^d$ used in [8] is a simplification of a suggestion of Drinfeld's (the latter potentially works for an arbitrary surface $\boldsymbol{S}$). We are trying to express points of $\mathcal{U}_G^d$ (one may call them quasi-bundles) by replacing the original problem for the surface $\mathbb{P}^2$ by another problem for the curve $\mathbb{P}^1$. Let us first generalize the problem to the case of G-bundles with a parabolic structure along a fixed straight line.

Namely, let $\boldsymbol{S} = \mathbb{P}^2$ and let $\mathbb{P}^1_\infty \subset \boldsymbol{S}$ be the "infinite line" (so that $\boldsymbol{S} \backslash \mathbb{P}^1_\infty = \mathbb{C}^2$). Let also Let $C \simeq \mathbb{P}^1 \subset \boldsymbol{S}$ denote the horizontal line in $\boldsymbol{S}$. Choose a parabolic subgroup $P \subset G$. Let $\mathrm{Bun}_{G,P}$ denote the moduli space of the following objects:

1) A principal G-bundle $\mathcal{H}_G$ on $\boldsymbol{S}$;

2) A trivialization of $\mathcal{H}_G$ on $\mathbb{P}^1_\infty \subset \boldsymbol{S}$;

3) A reduction of $\mathcal{H}_G$ to P on C compatible with the trivialization of $\mathcal{H}_G$ on C.

Let us describe the connected components of $\mathrm{Bun}_{G,P}$. Let M be the Levi group of P. Denote by $\check{M}$ the *Langlands dual* group of M and let $Z(\check{M})$ be its center. We denote by $\Lambda_{G,P}$ the lattice of characters of $Z(\check{M})$. Let also $\Lambda^{\mathrm{aff}}_{G,P} = \Lambda_{G,P} \times \mathbb{Z}$ be the lattice of characters of $Z(\check{M}) \times \mathbb{C}^*$. Note that $\Lambda^{\mathrm{aff}}_{G,G} = \mathbb{Z}$.

The lattice $\Lambda^{\mathrm{aff}}_{G,P}$ contains canonical semi-group $\Lambda^{\mathrm{aff},+}_{G,P}$ of positive elements. It is not difficult to see that the connected components of $\mathrm{Bun}_{G,P}$ are parameterized by the elements of $\Lambda^{\mathrm{aff},+}_{G,P}$:

$$\mathrm{Bun}_{G,P} = \bigcup_{\theta_{\mathrm{aff}} \in \Lambda^{\mathrm{aff},+}_{G,P}} \mathrm{Bun}^{\theta_{\mathrm{aff}}}_{G,P}.$$

Typically, for $\theta_{\mathrm{aff}} \in \Lambda^{\mathrm{aff}}_{G,P}$ we shall write $\theta_{\mathrm{aff}} = (\theta, d)$ where $\theta \in \Lambda_{G,P}$ and $d \in \mathbb{Z}$.

One would also like to construct the corresponding "Uhlenbeck scheme" $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$ stratified in the following way (the reader should compare it with (3.2)):

$$\mathcal{U}^{\theta}_{G,P} = \bigcup_{\mu_{\mathrm{aff}} \in \Lambda^{\mathrm{aff},+}_{G,P}} \mathrm{Bun}^{\theta_{\mathrm{aff}} - \mu_{\mathrm{aff}}}_{G,P} \times \mathrm{Sym}^{\mu_{\mathrm{aff}}}(\mathbb{C}^2). \tag{5.2}$$

The idea of the construction is as follows. Let us consider the scheme classifying triples $(\mathcal{F}_G, \beta, \gamma)$, where

1) $\mathcal{F}_G$ is a principal G-bundle on $\mathbb{P}^1$;

2) β is a trivialization of $\mathcal{F}_G$ on the formal neighborhood of $\infty \in \mathbb{P}^1$;

3) γ is a reduction to P of the fiber of $\mathcal{F}_G$ at $0 \in \mathbb{P}^1$.

It is easy to see that this scheme is canonically isomorphic to the thick partial flag variety $\boldsymbol{X}^{\mathrm{aff}}_{G,P} = G(\mathcal{K})/\boldsymbol{I}_P$. Under this identification the point $e^{\mathrm{aff}}_{G,P} \in \boldsymbol{X}^{\mathrm{aff}}_{G,P}$ corresponding to the unit element of G corresponds to the trivial $\mathcal{F}_G$ with the trivial trivialization.

It is explained in [8] that the variety $\mathrm{Bun}_{G,P}$ is canonically isomorphic to the scheme classifying *based maps* from (C, ∞_C) to $(\boldsymbol{X}^{\mathrm{aff}}_{G,P}, e^{\mathrm{aff}}_{G,P})$ (i.e. maps from C to $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ sending ∞_C to $e^{\mathrm{aff}}_{G,P}$).

One of the main results of [8] gives an explicit description of the Intersection Cohomology sheaf of all $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$. We shall not reproduce the full answer here; we shall only say that this answer is formulated in terms of the Lie algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$ – the affine Lie algebra whose Dynkin diagram is dual to that of $\mathfrak{g}_{\mathrm{aff}}$. Note that in general $\check{\mathfrak{g}}_{\mathrm{aff}} \neq (\check{\mathfrak{g}})_{\mathrm{aff}}$; in fact $\check{\mathfrak{g}}_{\mathrm{aff}}$ may result to be a twisted affine Lie algebra (thus it is not isomorphic to the affinization of any finite-dimensional $\mathfrak{g}$). We regard it as one of the first glimpses to (not yet formulated) Langlands duality for affine Lie algebras.

The proof of our computation of the IC-sheaves is also of independent interest. Namely, since in the affine case the results of [37] are not available we must have a different way to see the algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$ from the above geometry. In [8] we first do on the combinatorial level; namely we realize the canonical *Kashiwara crystal* discussed in [28] in terms of the varieties $\mathrm{Bun}^{\theta_{\mathrm{aff}}}_{G,B}$ (the idea of this realization is based on the earlier work [11]). We then use this geometric construction of crystals to compute the IC-sheaves of $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,B}$ (the answer is very similar to Theorem 3.1) which subsequently allows us to do it also for all $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$ using techniques similar to those developed in [9] (in particular, we compute the IC-sheaf for $P = G$ which is probably the most interesting case).

5.3. The universal Verma module for $\check{\mathfrak{g}}_{\mathrm{aff}}$. The scheme $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,B}$ is endowed with a natural action of $T \times (\mathbb{C}^*)^2$ (here $T \subset G$ acts by changing the trivialization of $\mathcal{H}_G$ (cf. the previous subsection) at $\mathbb{P}^1_\infty$ and $(\mathbb{C}^*)^2$ acts on $\boldsymbol{S} = \mathbb{P}^2$ preserving $\mathbb{P}^1_\infty$ and C). Note that the field $\mathcal{K}_{T\times(\mathbb{C}^*)^2}$ can be thought of as a field of rational functions of the variables $a \in \mathfrak{t}$, $\varepsilon_1, \varepsilon_2 \in \mathbb{C}$. Define

$$\mathrm{IH}^{\theta_{\mathrm{aff}}}_{G,B} = \mathrm{IH}^*_{T\times(\mathbb{C}^*)^2}(\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,B}) \underset{\mathcal{A}_{T\times(\mathbb{C}^*)^2}}{\otimes} \mathcal{K}_{T\times(\mathbb{C}^*)^2}, \quad \mathrm{IH}^{\mathrm{aff}}_{G,B} = \bigoplus_{\theta\in\Lambda^{\mathrm{aff},+}_{G,B}} \mathrm{IH}^{\theta_{\mathrm{aff}}}_{G,B}.$$

Thus $\mathrm{IH}^{\theta_{\mathrm{aff}}}_{G,B}$ is a vector space over $\mathcal{K}_{T\times(\mathbb{C}^*)^2}$ which is endowed with an intersection pairing $\langle\cdot,\cdot\rangle_{\theta_{\mathrm{aff}}}$ taking values in $\mathcal{K}_{T\times(\mathbb{C}^*)^2}$. Also, for each θ_{aff} as above we have the canonical element $1^{\theta_{\mathrm{aff}}}_{G,B} \in \mathrm{IH}^{\theta_{\mathrm{aff}}}_{G,B}$ corresponding to the unit cohomology class.

In [5] we show that the Lie algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$ acts naturally on $\mathrm{IH}^{\theta_{\mathrm{aff}}}_{G,B}$; the corresponding $\check{\mathfrak{g}}_{\mathrm{aff}}$-module is naturally isomorphic to the Verma module $M(\lambda_{\mathrm{aff}})$ where $\lambda_{\mathrm{aff}} = \frac{(a,\varepsilon_2)}{\varepsilon_1} + \rho_{\mathrm{aff}}$ (cf. [5] for more details). Note that this is very similar to statement 1) from Subsection 3.5; we also have analogs of the statements 2), 3), 4).

6. Applications to gauge theory and quantum cohomology of (affine) flag manifolds

6.1. The partition function. Recall that in the previous section we considered the moduli space Bun_G^d of G-bundles on $\boldsymbol{S} = \mathbb{P}^2$ trivialized at $\mathbb{P}^1_\infty \subset \boldsymbol{S}$ and having 2nd Chern class equal to d. We also have the scheme $\mathcal{U}_G^d$ containing Bun_G^d as an open subset. The group $G \times \mathrm{GL}(2)$ acts naturally on $\mathcal{U}_G^d$ where G acts by changing the trivialization at $\mathbb{P}^1_\infty$ and $\mathrm{GL}(2)$ acts on $\boldsymbol{S}$ preserving $\mathbb{P}^1_\infty$.

Thus we may consider (cf. [5] [41], [40] for precise definitions) the *equivariant integral*

$$\int_{\mathcal{U}_G^d} 1^d$$

of the unit $G \times \mathrm{GL}(2)$-equivariant cohomology class (which we denote by 1^d) over $\mathcal{U}_G^d$; the integral takes values in the field $\mathcal{K}$ which is the field of fractions of the algebra $\mathcal{A} = H^*_{G\times \mathrm{GL}(2)}(pt)$. Note that $\mathcal{A}$ is canonically isomorphic to the algebra of polynomial functions on the Lie algebra $\mathfrak{g} \times \mathfrak{gl}(2)$ which are invariant with respect to the adjoint action. Thus each $\int_{\mathcal{U}_G^d} 1^d$ may naturally be regarded as a rational function of $a \in \mathfrak{t}$ and $(\varepsilon_1, \varepsilon_2) \in \mathbb{C}^2$; this function must be invariant with respect to the natural action of W on $\mathfrak{t}$ and with respect to interchanging ε_1 and ε_2.

Consider now the generating function

$$\mathcal{Z} = \sum_{d=0}^{\infty} Q^d \int_{\mathcal{U}_G^d} 1^d.$$

It can (and should) be thought of as a function of the variables $\mathfrak{q}$ and $a, \varepsilon_1, \varepsilon_2$ as before. In [41] it was conjectured that the first term of the asymptotic in the limit $\lim_{\varepsilon_1,\varepsilon_2\to 0} \ln \mathcal{Z}$ is closely related to *Seiberg–Witten prepotential* of G.[4] This can be thought of as a rigorous mathematical formulation of the results of Seiberg and Witten from 1994. For $G = \mathrm{SL}(n)$ this conjecture has been proved in [42] and [40]. Also in [41] an explicit combinatorial expression for $\mathcal{Z}$ has been found. We are going to give a sketch of the proof of this conjecture for arbitrary G.

6.2. Parabolic generalization of the partition function. Recall from the previous section that for any parabolic $P \subset G$ we have the varieties $\mathrm{Bun}^{\theta_{\mathrm{aff}}}_{G,P}$ and $\mathcal{U}^{\theta}_{G,P}$. The latter contains the former as a dense open subset. Is is easy to see that $\mathcal{U}^{\theta}_{G,P}$ has a natural action of the group $M \times (\mathbb{C}^*)^2$ where M denotes the Levi subgroup of P. Also

[4] In fact, in [41] this conjecture is only formulated for $G = \mathrm{SL}(n)$ but the generalization to other groups is pretty straightforward. Also, one can reformulate everything in the language of moduli spaces of anti-selfdual connections of the 4-sphere S^4 rather than in terms of holomorphic (= algebraic) G-bundles on $\mathbb{P}^2$; this, perhaps, is closer to the physical origins of the problem.

we have the field $\mathcal{K}_{M\times(\mathbb{C}^*)^2}$ which is isomorphic to the field of rational functions on $\mathfrak{m}\times\mathbb{C}^2$ which are invariant with respect to the adjoint action.

Let $T\subset M$ be a maximal torus. Then one can show that $(\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P})^{T\times(\mathbb{C}^*)^2}$ consists of one point. This guarantees that we may consider the integral $\int_{\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}} 1^{\theta_{\mathrm{aff}}}_{G,P}$ where $1^{\theta_{\mathrm{aff}}}_{G,P}$ denotes the unit class in $H^*_{M\times(\mathbb{C}^*)^2}(\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P},\mathbb{C})$. The result can be thought of as a rational function on $\mathfrak{m}\times\mathbb{C}^2$ which is invariant with respect to the adjoint action of M. Define

$$\mathcal{Z}^{\mathrm{aff}}_{G,P}=\sum_{\theta\in\Lambda^{\mathrm{aff}}_{G,P}}\mathfrak{q}^{\theta_{\mathrm{aff}}}_{\mathrm{aff}}\int\limits_{\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}}1^{\theta_{\mathrm{aff}}}_{G,P}. \tag{6.1}$$

One should think of $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ as a formal power series in $\mathfrak{q}_{\mathrm{aff}}\in Z(\check{M})\times\mathbb{C}^*$ with values in the space of ad-invariant rational functions on $\mathfrak{m}\times\mathbb{C}^2$. Typically, we shall write $\mathfrak{q}_{\mathrm{aff}}=(\mathfrak{q},Q)$ where $\mathfrak{q}\in Z(\check{M})$ and $Q\in\mathbb{C}^*$. Also we shall denote an element of $\mathfrak{m}\times\mathbb{C}^2$ by $(a,\varepsilon_1,\varepsilon_2)$ or (sometimes it will be more convenient) by $(a,\hbar,\varepsilon)$ (note that for general P (unlike in the case $P=G$) the function $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ is not symmetric with respect to switching ε_1 and ε_2).

6.3. The "finite-dimensional" analog. Recall that the space $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$ is closely related to the space of based quasi-maps $C=\mathbb{P}^1\to\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ of degree θ_{aff}. Since the scheme $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$ may (and should) be thought of as a partial flag variety for $\mathfrak{g}_{\mathrm{aff}}$ it is natural to consider the following "finite-dimensional" analog of the above problem. Recall that we denote by ${}^{\mathrm{b}}\mathcal{M}^{\theta}_{G,P}$ the moduli space of based maps from (C,∞_C) to $(X_{G,P},e_{G,P})$ of degree θ, i.e. the moduli space of maps $C\to X_{G,P}$ which send ∞_C to $e_{G,P}$. This space is acted on by the group $M\times\mathbb{C}^*$.

We now introduce the "finite-dimensional" analog of the partition function (6.1). As before let

$$\mathcal{A}_{M\times\mathbb{C}^*}=H^*_{M\times\mathbb{C}^*}(pt,\mathbb{C})$$

and denote by $\mathcal{K}_{M\times\mathbb{C}^*}$ its field of fractions. Let also $1^{\theta}_{G,P}$ denote the unit class in the $M\times\mathbb{C}^*$-equivariant cohomology of ${}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,P}$. Then we define

$$\mathcal{Z}_{G,P}=\sum_{\theta\in\Lambda^{\theta}_{G,P}}\mathfrak{q}^{\theta}\int\limits_{{}^{\mathrm{b}}\mathcal{QM}^{\theta}_{G,P}}1^{\theta}_{G,P}. \tag{6.2}$$

This is a formal series in $q\in Z(\check{M})$ with values in the field $\mathcal{K}_{M\times\mathbb{C}^*}$ of M-invariant rational functions on $\mathfrak{m}\times\mathbb{C}$.

In fact the function $\mathcal{Z}_{G,P}$ is a familiar object in Gromov–Witten theory: in [5] we show that up to a simple factor $\mathcal{Z}_{G,P}$ is the so-called *equivariant J-function* of $X_{G,P}$; this function is in some sense responsible for the (small) quantum cohomology of $X_{G,P}$. Thus the problem of computation of the function $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ may be

thought of as the problem of computation of the (not yet rigorously defined) quantum cohomology of the affine flag manifolds $\boldsymbol{X}^{\mathrm{aff}}_{G,P}$. Note that in the case $G = \mathrm{SL}(n)$ and $P = B$ a heuristic computation of the latter ring in terms of the so-called *periodic Toda lattice* was done in [25]; our results discussed presented in the next subsection are compatible with this computation.

6.4. Computation of the partition functions in the Borel case. We believe that it should be possible to express the function $\mathcal{Z}_{G,P}$ (resp. the function $\mathcal{Z}^{\mathrm{aff}}_{G,P}$) in terms of representation theory of the Lie algebra $\check{\mathfrak{g}}$ (resp. $\check{\mathfrak{g}}_{\mathrm{aff}}$) – by the definition this is a Lie algebra whose root system is dual to that of $\mathfrak{g}$ (resp. to that of $\mathfrak{g}_{\mathrm{aff}}$). One of the main results of [5] gives such a calculation of the functions $\mathcal{Z}_{G,B}$ and $\mathcal{Z}^{\mathrm{aff}}_{G,B}$ where $B \subset G$ is a Borel subgroup of G. Roughly speaking we show that $\mathcal{Z}_{G,B}$ (resp. $\mathcal{Z}^{\mathrm{aff}}_{G,B}$) is equal to *Whittaker matrix coefficient* of the Verma module over $\check{\mathfrak{g}}$ (resp. over $\check{\mathfrak{g}}_{\mathrm{aff}}$) whose lowest weight given by $\frac{a}{\hbar}+\rho$ (resp. $\frac{(a,\varepsilon_1)}{\varepsilon_2}+\rho_{\mathrm{aff}}$ where a, $\hbar$, ε_1 and ε_2 are as in Subsection 5.3 (here we regard (a,ε_1) as a weight for the dual affine algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$; this is explained carefully in Section 3 of [5]). These statements in fact follow immediately from the results of Subsections 3.5 and 5.3 after one formulates the definition of equivariant integration using intersection cohomology (this is done is [5]).

The above description of the partition function allows one to produce certain differential equations which are satisfied by the functions $\mathcal{Z}_{G,B}$ and $\mathcal{Z}^{\mathrm{aff}}_{G,B}$. More precisely, we show that the function $\mathfrak{q}^{\frac{a}{\hbar}}\mathcal{Z}_{G,B}$ is an eigen-function of the *quantum Toda hamiltonians* associated with $\check{\mathfrak{g}}$ with eigen-values determined (in the natural way) by a (we refer the reader to [13] for the definition of (affine) Toda integrable system and its relation with Whittaker functions). In this way we reprove the results of [24] and [31] about (equivariant) quantum cohomology of the flag varieties $X_{G,P}$. In the affine case one can also show that $\mathfrak{q}^{\frac{a}{\hbar}}\mathcal{Z}^{\mathrm{aff}}_{G,B}$ is an eigen-function of a certain differential operator which has order 2 ("non-stationary analog" of the affine quadratic Toda hamiltonian). In [6] we explain how this allows to compute the asymptotics of *all* the functions $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ when $\varepsilon_1, \varepsilon_2 \to 0$ in another publication. We also show in [6] that this implies the Nekrasov conjecture mentioned above for arbitrary G.

7. Some open problems

In this section we present a list of open problems that related to the subjects in the preceding sections of the paper.

7.1. Normality and rational singularities. We conjecture that the schemes $\mathcal{QM}^{\theta}_{G,P}$ and $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$ are normal and have rational singularities. From purely technical point of view one needs to know this in order to attack problem 2. On the other hand, this statement seems to be important for the following reason. It is explained in [9] and [14] that one should think about the singularities of the schemes $\mathcal{QM}^{\theta}_{G,P}$ as a

finite-dimensional model for the singularities of the so-called *semi-infinite Schubert varieties* (cf. [14] for a more detailed discussion of this). On the other hand, one knows (cf. [13] and references therein) that usual (both finite and affine) Schubert varieties *are* normal and have rational singularities. Hence in the case of $\mathcal{QM}^{\theta}_{G,P}$ our conjectures can be thought of as a generalization of this result to the semi-infinite Schubert varieties. In some special case these conjectures were proved in [43].

7.2. Computation of parabolic partition functions. It would be very interesting to express the functions $\mathcal{Z}_{G,P}$ (resp. $\mathcal{Z}^{\mathrm{aff}}_{G,P}$) in terms of representation theory of the algebra $\check{\mathfrak{g}}$ (resp. $\check{\mathfrak{g}}_{\mathrm{aff}}$). Let us note that a combinatorial expression for all these functions in the case $G = \mathrm{SL}(n)$ was found in [34], but we do not know how to interpret this answer in terms of representation theory.

7.3. Other cohomology theories. The functions $\mathcal{Z}_{G,P}$ and $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ have their K-theoretic counterparts $\mathcal{Z}^{K}_{G,P}$ and $\mathcal{Z}^{K,\mathrm{aff}}_{G,P}$ (to define those one needs to replace the equivariant integrals considered in (6.1) and (6.2) by the corresponding integrals in equivariant K-theory). The function $\mathcal{Z}^{K}_{G,P}$ is exactly the K-theoretic J-function of $\mathcal{G}_{G,P}$ as defined in [22]. We would like to express these functions using the representation theory of the quantum group $U_q(\check{\mathfrak{g}})$ (resp. $U_q(\check{\mathfrak{g}}_{\mathrm{aff}})$). Presumably, in order to do this one should be able to construct geometrically some representations of these quantum groups. For $P = B$ and $G = \mathrm{SL}(n)$ this is done in [7]. For $P = G = \mathrm{SL}(n)$ the K-theoretic partition function $\mathcal{Z}^{K,\mathrm{aff}}_{G}$ was also studied in [28].

Let us also recall that in [6] the authors use the results described in Section 6.4 in order to connect certain asymptotic of the function $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ with the *Seiberg–Witten prepotential* corresponding to the classical Toda integrable system associated with the Lie algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$. We would like to generalize this to the K-theoretic partition functions. This should involve some interesting interplay between quantum affine algebras and certain difference equations (which should be thought of non-stationary deformations of known integrable difference equations of Toda type). The corresponding classical integrable system describing the asymptotic should be the so-called *relativistic Toda system* associated with the Lie algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$.

It would also be interesting to generalize this to the case when K-theory is replaced by any elliptic cohomology theory. Again, for $P = G = \mathrm{SL}(n)$ (in the affine case) this is done in [26].

7.4. Chern classes of the tangent bundle and the Calogero–Moser system. As was mentioned above the partitions functions $\mathcal{Z}^{\mathrm{aff}}_{G,P}$ are related to the pure $N = 2$ super-symmetric gauge theory in 4 dimensions. One should also have an extension of the above result for gauge theory *with matter*. This means that instead of considering equivariant integrals of the unit cohomology class we should consider integrals of the Chern classes of various natural bundles on the moduli spaces in question. For example in the case of *adjoint matter* one should integrate the Chern polynomial of the tangent bundle of $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,P}$ (this, of course, has to be properly interpreted since the

variety in question is singular). For $P = G = \mathrm{SL}(n)$ such functions are studied in [28] and the corresponding asymptotic of the partition function is shown there to be related with the prepotential of the classical *elliptic Calogero–Moser* integrable system. We expect that for $P = B$ (and for general G) the corresponding partition function should be closely related with the universal eigen-function of the corresponding non-stationary deformation of the quantum Calogero–Moser hamiltonian associated with the Lie algebra $\check{\mathfrak{g}}_{\mathrm{aff}}$. Similar statement should also hold in the finite case (i.e. when we integrate over $\mathcal{QM}^{\theta}_{G,B}$'s and not over $\mathcal{U}^{\theta_{\mathrm{aff}}}_{G,B}$'s). In particular, in the finite case we should get a geometric interpretation of the universal eigen-function of the quantum Calogero–Moser system associated with the Lie algebra $\check{\mathfrak{g}}$.

7.5. Functional equation for parabolic Eisenstein series. It will be very important to generalize the functional equation for geometric Eisenstein series discussed in Subsection 4.4 to the case of parabolic Eisenstein series. More precisely, given a parabolic subgroup $P \subset G$ with the Levi subgroup M in Section 4 we discussed the Eisenstein series functor Eis^G_M. In fact this notation is slightly misleading since this functor actually depends not just on M but also on P. Let us now use the notation $\mathrm{Eis}^G_{M,P}$ for it. With this notation the "functional equation" problem reads as follows: given two parabolic subgroups $P, Q \subset G$ containing *the same* Levi subgroup M, construct as isomorphism between the functors $\mathrm{Eis}^G_{M,P}$ and $\mathrm{Eis}^G_{M,Q}$ restricted to some large subcategory of "regular" sheaves inside $\mathrm{Sh}(M)$.

References

[1] Arkhipov, S., Braverman, A., Bezrukavnikov, R., Gaitsgory, D., Mirković, I., Modules over the small quantum group and semi-infinite flag manifold. *Transform. Groups* **10** (3–4) (2005), 279–362.

[2] Balaji, V., Principal bundles on projective varieties and the Donaldson-Uhlenbeck compactification, math.AG/0505106.

[3] Baranovsky, V., Ginzburg, V., Algebraic construction of the Uhlenbeck moduli space. Manuscript, 1998.

[4] Beilinson, A., Bernstein, J., Localisation de g-modules. *C. R. Acad. Sci. Paris Sér. I Math.* **292** (1) (1981), 15–18.

[5] Braverman, A., Instanton counting via affine Lie algebras I. In *Equivariant J-functions of (affine) flag manifolds and Whittaker vectors*, CRM Proc. Lecture Notes 38, Amer. Math. Soc., Providence, R.I., 2004, 113–132.

[6] Braverman, A., and Etingof, P., Instanton counting via affine Lie algebras II: from Whittaker vectors to the Seiberg-Witten prepotential. math.AG/0409441.

[7] Braverman, A., Finkelberg, M., Finite difference quantum Toda lattice via equivariant K-theory. *Transform. Groups* **10** (2005), 363–386.

[8] Braverman, A., Finkelberg, M., Gaitsgory, D., Uhlenbeck spaces via affine Lie algebras. In *The unity of mathematics* (In honor of the ninetieth birthday of I. M. Gelfand), Progr. Math. 244, Birkhäuser, Boston, MA, 2006, 17–135.

[9] Braverman, A., Finkelberg, M., Gaitsgory, D., Mirković, I., Intersection cohomology of Drinfeld's compactifications. *Selecta Math.* (*N.S.*) **8** (3) (2002), 381–418.

[10] Braverman, A., Gaitsgory, D., Geometric Eisenstein series. *Invent. Math.* **150** (2) (2002), 287–384.

[11] Braverman, A., Gaitsgory, D., Crystals via the affine Grassmannian. *Duke Math. J.* **107** (3) (2001), 561–575.

[12] Donaldson, S. K., Connections, cohomology and the intersection forms of four-manifolds. *J. Differential Geom.* **24** (1986), 275–341.

[13] Etingof, P., Whittaker functions on quantum groups and q-deformed Toda operators. In *Differential topology, infinite-dimensional Lie algebras, and applications*, Amer. Math. Soc. Transl. Ser. (2) 194, Amer. Math. Soc., Providence, R.I., 1999, 9–25.

[14] Faltings, G., Algebraic loop groups and moduli spaces of bundles. *J. Eur. Math. Soc.* (*JEMS*) **5** (1) (2003), 41–68.

[15] Feigin, B., Finkelberg, M., Kuznetsov, A., Mirković, I., Semi-infinite flags. II. Local and global intersection cohomology of quasimaps' spaces. In *Differential topology, infinite-dimensional Lie algebras, and applications*, Amer. Math. Soc. Transl. Ser. (2) 194, Amer. Math. Soc., Providence, R.I., 1999, 113–148.

[16] Finkelberg, M., Gaitsgory, D., Kuznetsov, A., Uhlenbeck spaces for $\mathbb{A}^2$ and affine Lie algebra $\widehat{sl}_n$. *Publ. Res. Inst. Math. Sci.* **39** (2003), 721–766.

[17] Finkelberg, M., Kuznetsov, A., Global Intersection Cohomology of Quasimaps' spaces. *Internat. Math. Res. Notices* **7** (1997), 301–328.

[18] Frenkel, E., Gaitsgory, D., and Vilonen, K., Whittaker patterns in the geometry of moduli spaces of bundles on curves. *Ann. of Math.* (2) **153** (3) (2001), 699–748.

[19] Frenkel, E., Gaitsgory, D., and Vilonen, K., On the geometric Langlands conjecture. *J. Amer. Math. Soc.* **15** (2) (2002), 367–417.

[20] Gaitsgory, D., On a vanishing conjecture appearing in the geometric Langlands correspondence, *Ann. of Math.* (2) **160** (2) (2004), 617–682.

[21] Gaitsgory, D., Geometric Langlands correspondence for GL_n. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002 571–582.

[22] Gaitsgory, D., On de Jong's conjecture. math.AG/0402184.

[23] Givental, A., Lee, Y.-P., Quantum K-theory of flag manifolds, finite-difference Toda lattices and quantum groups. *Invent. Math.* **151** (2003), 193–219.

[24] Givental, A., and Kim, B., Quantum cohomology of flag manifolds and Toda lattices. *Comm. Math. Phys.* **168** (1995), 609–641.

[25] Guest, M. A., Otofuji, T., Quantum cohomology and the periodic Toda lattice. *Comm. Math. Phys.* **217** (3) (2001), 475–487.

[26] Hollowood, T. J., Iqbal, A., Vafa, C., Matrix Models, Geometric Engineering and Elliptic Genera. hep-th/0310272.

[27] Kashiwara, M., Kazhdan-Lusztig conjecture for a symmetrizable Kac-Moody Lie algebra. In *The Grothendieck Festschrift*, Vol. II, Progr. Math. 87, Birkhäuser, Boston, MA, 1990, 407–433.

[28] Kashiwara, M., Saito, Y., Geometric construction of crystal bases, *Duke Math. J.* **89** (1) (1997), 9–36.

[29] Kazhdan, D., Lusztig, G., Representations of Coxeter groups and Hecke algebras. *Invent. Math.* **53** (2) (1979), 165–184.

[30] Kuznetsov, A., Laumon's resolution of Drinfeld's compactification is small. *Math. Res. Lett.* **4** (2–3) (1997), 349–364.

[31] Kim, B., Quantum cohomology of flag manifolds G/P and quantum Toda lattices. *Ann. of Math.* **149** (1999), 129–148.

[32] Laumon, G., Faisceaux Automorphes Liés aux Séries d'Eisenstein. *Perspect. Math.* **10** (1990), 227–281.

[33] J. Li, Algebraic geometric interpretation of Donaldson's polynomial invariants. *J. Differential Geom.* **37** (1993), 417–466.

[34] Lian, B. H., Liu, C.-H., Liu, K., Yau, S.-T., The S^1 fixed points in Quot-schemes and mirror principle computations. In *Vector bundles and representation theory* (Columbia, MO, 2002), Contemp. Math. 322, Amer. Math. Soc., Providence, R.I., 2003, 165–194.

[35] Lusztig, G., Singularities, character formulas, and a q-analog of weight multiplicities. *Astérisque* **101–102** (1983), 208–229.

[36] Lusztig, G., Periodic W-graphs. *Represent. Theory* **1** (1997), 207–279 (electronic).

[37] Mirković, I., Vilonen, K., Perverse sheaves on affine Grassmannians and Langlands duality. *Math. Res. Lett.* **7** (1) (2000), 13–24

[38] Morgan, J. W., Comparison of the Donaldson polynomial invariants with their algebro-geometric analogues. *Topology* **32** (1993), 449–488.

[39] Nakajima, H., *Lectures on Hilbert schemes of points on surfaces*. Univ. Lecture Ser. 18, Amer. Math. Soc., Providence, R.I., 1999.

[40] Nakajima, H., Yoshioka, K., Lectures on instanton counting. In *Algebraic structures and moduli spaces*, CRM Proc. Lecture Notes 38, Amer. Math. Soc., Providence, R.I., 2004, 31–101.

[41] Nekrasov, N., Seiberg-Witten prepotential from instanton counting. *Adv. Theor. Math. Phys.* **7** (5) (2003), 831–864.

[42] Nekrasov, N., Okoun'kov, A., Seiberg-Witten theory and random partitions. In *The unity of mathematics* (In honor of the ninetieth birthday of I. M. Gelfand), Progr. Math. 244, Birkhäuser, Boston, MA, 2006, 525–596.

[43] Sottile, F., Sturmfels, B., A sagbi basis for the quantum Grassmannian. *J. Pure Appl. Algebra* **158** (2–3) (2001), 347–366.

[44] Sevostyanov, A., Quantum deformation of Whittaker modules and the Toda lattice. *Duke Math. J.* **105** (2000), 211–238.

[45] Uhlenbeck, K., Connections with L^p bounds on curvature, *Comm. Math. Phys.* **83** (1982), 31–42.

Department of Mathematics, Brown University, Providence, RI, U.S.A.
E-mail: braval@math.brown.edu

On the local Langlands and Jacquet–Langlands correspondences

Guy Henniart

Abstract. Let F be a locally compact non Archimedean field, and D a division algebra with centre F and finite dimension d^2 over F. Fix an integer $r \geqslant 1$, and let $G = \mathrm{GL}_n(F)$, $G' = \mathrm{GL}_r(D)$, where $n = rd$. Smooth irreducible representations of G' are related to those of G via the Jacquet–Langlands correspondence, whereas the Langlands correspondence relates such representations of G to degree n representations of the absolute Galois group of F. We review some recent results on those correspondences, in particular on their explicit description.

Résumé. Soient F un corps commutatif localement compact non archimédien, et D un corps gauche de centre F et de dimension finie d^2 sur F. Fixons un entier $r \geqslant 1$ et posons $n = rd$, $G' = \mathrm{GL}_r(D)$, $G = \mathrm{GL}_n(F)$. Les représentations lisses irréductibles de G' sont reliées à celles de G par la correspondance de Jacquet–Langlands, tandis que la correspondance de Langlands relie celles de G aux représentations de dimension n du groupe de Galois absolu de F. Nous passons en revue quelques résultats récents concernant ces correspondances, et en particulier leur description explicite.

Mathematics Subject Classification (2000). Primary 22E50.

Keywords. Local field, smooth representations, Jacquet–Langlands correspondence, Langlands correspondence.

1. Smooth representations

Let F be a locally compact non-Archimedean field, and p its residue characteristic. So F is a finite extension of the field $\mathbb{Q}_p$ of p-adic numbers, or of the field of Laurent power series $\mathbb{F}_p((X))$ in one variable X.

If $\mathbb{H}$ is a reductive linear algebraic over F, then the group $H = \mathbb{H}(F)$, with its natural topology coming from F, is locally profinite: there is a basis of neighbourhoods of identity consisting of open compact subgroups. We shall be interested only in the following special cases. We let D be a division algebra with centre F and finite dimension over F; that dimension is necessarily a square d^2, $d \geqslant 1$. We fix an integer $r \geqslant 1$, put $n = rd$, and write G'_r or simply G' for $\mathrm{GL}_r(D)$ and G_n or simply G for $\mathrm{GL}_n(F)$. Both are examples of groups H as above, and G is the special case of G' where $d = 1$, $n = r$.

A representation of a locally profinite group H, on a complex vector space V, is *smooth* if every vector v in V has open stabilizer in H. With obvious morphisms smooth representations of H form an abelian category.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Simple objects in that category are called *irreducible*: a smooth representation (π, V) of H is irreducible if V is non-zero and contains no subspace invariant under H, other than $\{0\}$ and V. We write $\mathcal{A}(H)$ for the set of isomorphism classes of irreducible smooth representations of H.

A *character* of H is a group homomorphism into $\mathbb{C}^\times$, with open kernel. Characters of H parametrize isomorphism classes of smooth 1-dimensional representations.

We consider only locally profinite groups H such that for one – and hence for all – open compact subgroups K, H/K is *countable*. In that case Schur's lemma is valid for an irreducible smooth representation (π, V) of H, and in particular the centre $Z(H)$ of H acts on V via a character ω_π called the *central character* of π. In our case the centre of G' or G is made out of scalar matrices, and will be identified with $F^\times$.

2. Coefficients

If (π, V) is a smooth representation of a locally profinite group H, the group H acts on $V^* = \mathrm{Hom}(V, \mathbb{C})$ by $g \mapsto {}^t\pi(g^{-1})$. The subspace $V^\vee$ of V^* of linear functionals which are *smooth*, i.e. stabilized by an open subgroup of H, carries a smooth representation $(\pi^\vee, V^\vee)$ of V, called the *contragredient* of (π, V). The natural embedding $V \to V^{\vee\vee}$ is an isomorphism if and only if (π, V) is *admissible*, which means that for all open compact subgroups K of H, the subspace V^K of fixed points under K is finite-dimensional; in that case $(\pi^\vee, V^\vee)$ is also admissible. When H is reductive over F, a smooth representation of H of finite length is admissible.

A *coefficient* of (π, V) is a function c on H of the form $c(g) = \lambda(\pi(g)v)$ where $v \in V$, $\lambda \in V^\vee$. If (π, V) has a central character ω_π, then $c(zg) = \omega_\pi(z)c(g)$ for $z \in Z(H)$, $g \in H$.

Let (π, V) be *irreducible*. We say that it is *cuspidal* if the support of any coefficient is compact modulo $Z(H)$; we say it is *square integrable* (L^2) if ω_π is unitary and for any coefficient c of π, its absolute value $|c|$ is square integrable on $H/Z(H)$, for any Haar measure on that quotient group. We let $\mathcal{A}^o(H)$ (resp. $\mathcal{A}^2(H)$) be the subset of $\mathcal{A}(H)$ made out of cuspidal (resp. L^2) representations.

3. The Jacquet–Langlands correspondence

That correspondence is a special case of Langlands' functoriality principle, which very roughly speaking says that two groups sharing many conjugacy classes should share a large part of their representation theory. That is precisely the case for our groups G' and G.

An element g' of G' has a reduced characteristic polynomial $P(g')$, which is a monic polynomial of degree $n = rd$ in $F[T]$. We say that g' is regular semisimple if $P(g')$ has no repeated root in an algebraic closure of F, and that g' is regular elliptic

if moreover $P(g')$ is irreducible in $F[T]$. Two regular semisimple elements of G' are conjugate in G' if and only if they have the same characteristic polynomial. The set G'^s of regular semisimple elements in G' is open and dense, and the set G'^e of regular elliptic elements is open.

All that applies to G as a special case. For $g' \in G'^s$, there is $g \in G^s$, unique up to conjugation, such that $P(g) = P(g')$. For $g \in G^s$, there is a $g' \in G'^s$ such that $P(g) = P(g')$ if and only if all irreducible factors of $P(g)$ have degree a multiple of d; that is obviously the case for $g \in G^e$.

To express how G and G' can share some of their representation theory, recall that if we fix a Haar measure dg' on G', the convolution algebra $\mathcal{H}(G')$ of locally constant compactly supported functions on G' acts in any smooth representation (π', V') of G' via $\pi'(f)v' = \int_{G'} f(g')\pi(g')\,dg'$, for v' in V' and f in $\mathcal{H}(G')$. If (π', V') is admissible (in particular when π' has finite length), $\pi(f)$ has finite dimensional range hence has a trace.

Theorem 1. *Let (π', V') be a finite length smooth representation of G'. There is a unique locally constant function $\chi_{\pi'}$ on G'^s, locally integrable on G', such that for $f \in \mathcal{H}(G')$ we have* $\operatorname{tr} \pi'(f) = \int_{G'} \chi_{\pi'}(g') f(g')\,dg'$.

Note that by uniqueness $\chi_{\pi'}$ is invariant under conjugation. Also it does not depend on the choice of Haar measure dg' on G'. If f has support in G'^s, the result is due to Harish Chandra [27], [28], as is the general case when F has characteristic 0 [27]. When F has positive characteristic, it is more difficult and is due to B. Lemaire, first for G [46] and recently for G' [47].

Corollary. *Let $\pi'_1, \dots, \pi'_s$ be inequivalent smooth irreducible representations of G'. Then the functions $\chi_{\pi'_1}, \dots, \chi_{\pi'_s}$ are linearly independent.*

The Jacquet–Langlands correspondence is expressed by the following result.

Theorem 2. *There is a unique bijection $\pi \leftrightarrow \pi'$ between $\mathcal{A}^2(G)$ and $\mathcal{A}^2(G')$ such that $\chi_\pi(g) = (-1)^{n-r}\chi_{\pi'}(g')$ whenever $g \in G^e$, $g' \in G'^e$, $P(g) = P(g')$.*

When $n = 2$, where G' is $D^\times$, D the quaternion division algebra over F, the result is due to Jacquet and Langlands [39], hence the name. For $n = 3$ it is due to Flath [22] when F has characteristic 0 and to the author [33] in general. When $n > 3$, $r = 1$, it was established by Rogawski [50] in characteristic 0 and Badulescu [2] in positive characteristic.

Finally the general case $n > 3$, $r > 1$ was obtained by Deligne, Kazhdan and Vignéras [21] in characteristic 0 and Badulescu [3] in positive characteristic. Note that the method is global, it uses automorphic forms and trace formulas. Only when $n = 2$ a purely local proof is known [24], see also [14].

Remarks. 1) If $\pi \leftrightarrow \pi'$ as above, then $\omega_\pi = \omega_{\pi'}$ and $\pi^\vee \leftrightarrow \pi'^\vee$.

2) For π' a smooth representation of G' and χ a character of $F^\times$, we let $\chi\pi'$ be the smooth representation $g' \mapsto \chi \circ \operatorname{Nrd}(g')\pi'(g')$, where Nrd is the reduced norm

(the determinant in the special case of G). For $\pi \leftrightarrow \pi'$ as above, and χ unitary, then $\chi\pi \in \mathcal{A}^2(G)$, $\chi\pi' \in \mathcal{A}^2(G')$ and $\chi\pi \leftrightarrow \chi\pi'$.

4. Extending the Jacquet–Langlands correspondence

When $r > 1$, the groups G and G' share more conjugacy classes than just elliptic ones.

Theorem 3 ([4]). *If $\pi \in \mathcal{A}^2(G)$ and $\pi' \in \mathcal{A}^2(G')$ correspond as above, then*

$$\chi_\pi(g) = (-1)^{n-r}\chi_{\pi'}(g') \text{ whenever } g \in G^s,\ g' \in G'^s,\ P(g) = P(g').$$

It is natural to expect that G and G' also share more of their representation theory than just discrete series.

From Remark 2 of § 3, we readily extend the Jacquet–Langlands correspondence a bit, as follows.

A smooth representation π' of G' is called *essentially* L^2 if it is of the form $\chi\pi'_0$, where π'_0 is L^2 and χ is a character of $F^\times$. Writing $\mathcal{A}^d(G')$ for the essentially L^2 elements in $\mathcal{A}(G')$, we get a unique bijection $\pi \leftrightarrow \pi'$ between $\mathcal{A}^d(G)$ and $\mathcal{A}^d(G')$, for which the assertion of Theorem 3 is still valid, and then $\chi\pi \leftrightarrow \chi\pi'$ for all characters χ of $F^\times$, if $\pi \leftrightarrow \pi'$.

To go further, we have to use parabolic induction. For $i = 1, 2$, let r_i be a positive integer, and (π'_i, V_i) a smooth representation of G'_{r_i}. We consider the space $\mathcal{F}$ of functions $f : G'_{r_1+r_2} \to V_1 \otimes V_2$ which satisfy

$$f(mu\ g) = \delta(m)^{1/2}(\pi'_1(g_1) \otimes \pi'_2(g_2))(f(g))$$

for any $g \in G'_{r_1+r_2}$, any diagonal block matrix m in $G'_{r_1+r_2}$ with diagonal blocks $g_1 \in G'_{r_1}$ and $g_2 \in G'_{r_2}$, and any upper unipotent block matrix u with diagonal blocks 1_{r_1} and 1_{r_2}. Here δ is a specific character of $G'_{r_1} \times G'_{r_2}$ with positive real values, called the *modulus character*. The group $G'_{r_1+r_2}$ acts on $\mathcal{F}$ by right translations and the subspace $\mathcal{F}^\infty$ of functions stabilized by an open subgroup of $G'_{r_1+r_2}$ carries a smooth representation $i(\pi'_1, \pi'_2)$ of $G'_{r_1+r_2}$.

Note. The rôle of the modulus character is that if π'_1 and π'_2 are unitary – i.e. there is a G'_{r_i}-invariant hermitian positive definite form on the space V_i – then so is $i(\pi'_1, \pi'_2)$.

If π'_1 and π'_2 have finite length, the same is true of $i(\pi'_1, \pi'_2)$. Writing $\mathcal{R}'_r$ for the Grothendieck group of the category of finite length smooth representations of G'_r – this is a free $\mathbb{Z}$-module with basis $\mathcal{A}(G'_r)$ – we get a $\mathbb{Z}$-linear multiplication $\mathcal{R}'_{r_1} \times \mathcal{R}'_{r_2} \to \mathcal{R}'_{r_1+r_2}$ such that $[\pi'_1] \times [\pi'_2] = [i(\pi'_1, \pi'_2]$ where $[\pi']$ is the class in the Grothendieck group of the smooth representation π' of finite length. Putting $\mathcal{R}' = \bigoplus_{r \geqslant 1} \mathcal{R}'_r$, we get a graded ring ([61] for $d = 1$, [58]) which is *commutative*; as a special we get a graded ring $\mathcal{R} = \bigoplus_{k \geqslant 1} \mathcal{R}_k$ for the groups G_k.

Theorem (Badulescu, [4]). a) *There is a unique $\mathbb{Z}$-linear map $JL\colon \mathcal{R} \to \mathcal{R}'$ sending $\mathcal{R}_k$ to 0 if $d \nmid k$ and to $\mathcal{R}'_{k/d}$ if $d \mid k$, such that for $\pi \in \mathcal{R}$ and $\pi' = JL(\pi)$ we have*

$$\chi_\pi(g) = (-1)^{n-r}\chi_{\pi'}(g') \text{ whenever } g \in G^s_{rd},\ g' \in G'^s_r,\ P(g) = P(g').$$

b) *The map JL is the unique ring homomorphism which is trivial on $\mathcal{R}_k$ when $d \nmid k$ and extends the Jacquet–Langlands correspondence on $\mathcal{A}^d(G_{rd})$ for any integer $r \geqslant 1$.*

In fact $\mathcal{R}$ and $\mathcal{R}'$ are even Hopf algebras with a canonical involution, the Aubert–Bernstein–Zelevinsky involution [1], [61], and JL is a Hopf-algebra homomorphism, which preserves the involutions, up to predictable signs. The kernel of JL is the ideal generated by the $\mathcal{R}_k$ for $d \nmid k$.

It is expected that when π is a *unitary* smooth irreducible representation of G_k, then $JL([\pi])$ is either 0 – in particular if $d \nmid k$ – or plus or minus the class of a unitary smooth *irreducible* representation of $G'_{k/d}$. Badulescu [5] has proved that when π is a local component of some discrete series representation of $\mathrm{GL}_k(\mathbb{A}_E)$, where E is a number field with completion F at some finite place.

Remark. The unitary elements in $\mathcal{A}(G_k)$ have been classified by Tadić [57], and a similar classification is expected for $\mathcal{A}(G'_r)$ [58]. By recent work of Tadić, Badulescu and Renard [4], [5], [6], it is enough to prove the following conjecture, a result due to Bernstein [7] for G_k.

Conjecture. Let $\pi_i \in \mathcal{A}(G'_{r_i})$, $i = 1, 2$, be unitary. Then $\pi_1 \times \pi_2$ is irreducible.

5. The Langlands correspondence

Let $\overline{F}$ be a separable algebraic closure of F. The group $\mathrm{Gal}(\overline{F}/F)$ is then profinite, and class field theory gives a canonical bijection between characters of $\mathrm{Gal}(\overline{F}/F)$ and finite order characters of $F^\times$. Replacing $\mathrm{Gal}(\overline{F}/F)$ by a variant, the Weil group W_F [60, Appendix II], even gives a bijection between characters of W_F and characters of $F^\times = \mathrm{GL}_1(F)$.

It was Langlands' fundamental intuition that irreducible smooth representations of G_n, $n \geqslant 2$, are intimately related to degree n smooth representations of W_F, cuspidal representations of G_n corresponding to irreducible representations of W_F. Write $\mathcal{G}^o(n)$ for the set of isomorphism classes of irreducible degree n smooth representations of W_F.

Theorem. *There is a unique family of bijective maps $\mathcal{G}^o(n) \to \mathcal{A}^o(G_n)$, $\sigma \mapsto \pi(\sigma)$, such that*

1) *for $n = 1$, it is given by class field theory;*

2) *for $n, m \geqslant 1$, $\sigma \in \mathcal{G}^o(n)$, $\tau \in \mathcal{G}^o(m)$ we have*

$$L(\sigma \otimes \tau, s) = L(\pi(\sigma) \times \pi(\tau), s) \quad \text{and} \quad \varepsilon(\sigma \otimes \tau, s, \psi) = \varepsilon(\pi(\sigma) \times \pi(\tau), s, \psi)$$

for all non-trivial characters ψ of F.

Here the L-factors are of the form $P(p^{-s})^{-1}$, $P \in \mathbb{C}[X]$, $P(0) = 1$, and the ε factors are monomials in p^{-s}. The L-factor on the left is Artin's, whereas the ε-factors on the left have been defined by Langlands and Deligne [20], generalizing Tate's thesis [59] which concerns characters of $F^\times$, i.e. one dimensional representations of W_F. The L-factors and ε-factors on the left are those defined by Jacquet, Piatetski–Shapiro, and Shalika [40], or equivalently Shahidi [54], [55], a very different generalization of Tate's thesis. It is wonderfully ironic that both generalizations turn out to be "the same".

Remark. The central character of $\pi(\sigma)$, for $\sigma \in \mathcal{G}^o(n)$, corresponds to the character $\det \sigma$ of W_F via class field theory. Also $\pi(\sigma^\vee) = \pi(\sigma)^\vee$.

The theorem is due to Laumon, Rapoport and Stuhler [45] when F has positive characteristic, and to M. Harris and R. Taylor [32], see also [30], [31] for p-adic fields. Both proofs are global and geometric, using automorphic forms, trace formulas and the geometry of Drinfeld or Shimura modular varieties. They also rely on counting arguments due to the author [34].

The proof in positive characteristic is easier because one can exploit then the functional equation for the L function of an ℓ-adic representation of global Galois groups. Of course the theorem is now also a consequence of L. Lafforgue's result [43] establishing the Langlands conjecture for global fields of positive characteristic. In characteristic zero, the set of Galois L-functions for which one can prove a functional equation is much more restricted, and one has to use Harris' approach [29]. The method in [32] uses heavily the geometry of Shimura varieties at places of bad reduction, which has the advantage of yielding a geometric model for the Langlands correspondence – that is also the case for [45]. The author has given a simpler proof [36], which uses the same Shimura varieties but only at places of good reduction, where it is much easier to get – but one does not get the geometric model that way.

It is highly desirable to find a proof not relying on geometry, and ideally a purely local proof. However the program of C. J. Bushnell and the author [9 and references therein] faces obstacles in establishing the necessary properties of the ε-factors for pairs above. In the simplest cases $n = 2, 3$ the following can be proved ([41] for $n = 2$, [33] for $n = 3$) without geometry.

Theorem. *Let n be 2 or 3. There is a unique bijective map $\mathcal{G}^o(n) \to \mathcal{A}^o(G_n)$, $\sigma \mapsto \pi(\sigma)$, such that $\varepsilon(\chi\sigma, s, \psi) = \varepsilon(\chi\pi(\sigma), s, \psi)$ for all non-trivial additive characters ψ of F and all characters χ of $F^\times$.*

Here $\chi\sigma$ denotes the representation $g \mapsto \tilde{\chi}(g)\sigma(g)$ of W_F, where $\tilde{\chi}$ is the character of W_F corresponding to χ via class field theory.

The proof uses automorphic forms on global fields and trace formulas, even for $n = 2$. It is only recently that Bushnell and the author found a proof for $n = 2$, essentially local, which *does not* use *automorphic forms* [14].

Remark. The Langlands correspondence preserves more than the L and ε factors considered in the theorem. For example, the author shows in [37] that if $\sigma \in \mathcal{G}^o(n)$

then

$$L(\Lambda^2\sigma, s) = L(\tau(\sigma), \Lambda^2, s) \text{ and}$$
$$L(S^2\sigma, s) = L(\tau(\sigma), S^2, s)$$

where the L-factors on the right are those defined by Shahidi [55]. The corresponding ε-factors are also preserved, at least up to a non-zero constant (depending on σ).

6. Explicit Langlands correspondence in the tame case

All elements of $\mathcal{A}^o(G_n)$ have been constructed explicitly by Bushnell and Kutzko [15], see also [8]: for each $\pi \in \mathcal{A}^o(G_n)$ they describe an open subgroup J of G_n, which contains the centre Z and is compact mod Z, and a finite dimensional smooth irreducible representation λ of J such that π is obtained from λ by smooth induction (a simpler variant of the construction in § 4). When n is prime to p, the so-called *tame case*, the construction goes back to R. Howe [38], and when $n = p$ to Carayol [19], but the general case is much harder.

It is natural to ask for an explicit description of the Langlands correspondence in terms of such a construction. When p divides n it is out of reach at present: only the case $n = p = 2$ is reasonably settled [42], [14]. In the tame case Howe parametrized both $\mathcal{G}^o(n)$ and $\mathcal{A}^o(G_n)$ in terms of *admissible pairs* $(E/F, \theta)$, where E/F is a degree n extension, θ a character of $E^\times$ not factorizing through an intermediate norm $N_{F/E'}$ $F \subset E' \subset E$, $E' \neq E$, and such that if θ restricted to principal units of E does factorize, then E/E' is unramified. There is then a canonical map $(E/F, \theta) \mapsto \pi(E/F, \theta)$ giving a bijection between admissible pairs up to isomorphism and $\mathcal{A}^o(G_n)$; see [49] for a precise construction, or [12].

On the other hand, if $(E/F, \theta)$ is an admissible pair, θ can be seen as a character of the Weil group W_E, which is an open subgroup of index n of the Weil group W_F. We can then form $\sigma(E/F, \theta)$, the degree n smooth representation of W_F induced from the character θ of W_E. The map $(E/F, \theta) \mapsto \sigma(E/F, \theta)$ gives a bijection between admissible pairs, up to isomorphism, and $\mathcal{G}^o(n)$.

However, the determinant of $\sigma(E/F, \theta)$, seen as a character of $F^\times$, differs in general from the central character of $\pi(E/F, \theta)$, so $\pi(E/F, \theta)$ cannot be $\pi(\sigma(E/F, \theta))$. Recently C.J. Bushnell and the author obtained the following result, in the more general *essentially tame* situation where n is not necessarily prime to p but E/F is still tamely ramified of degree n.

Theorem ([12]). *Let $(E/F, \theta)$ be an admissible pair with E/F tamely ramified of degree n. Then there is a unique tamely ramified character $\mu = \mu(E/F, \theta)$ of $E^\times$ such that $\pi(\sigma(E/F, \theta)) = \pi(E/F, \mu\theta)$.*

The twisting character μ has been computed, at least when E/F is totally ramified [13]. The answer is easy to state only when n is odd; moreover when n is even, the answer *does not* coincide in general with the recipe conjectured in [49].

7. Explicit Jacquet–Langlands correspondence, $r = 1$

An explicit construction of $\mathcal{A}(G'_r)$ or even its cuspidal part $\mathcal{A}^o(G'_r)$, is not known in general (see § 8). However, when $r = 1$, the group $G'_1 = D^\times$ is compact modulo its centre Z and all its irreducible smooth representations are cuspidal and finite-dimensional. They have been constructed by E.-W. Zink [62] and, in terms closer to [15], by P. Broussous [16]. In the tame case, when d is prime to p, the construction goes back to R. Howe [38], see also [48], [49]: there is a natural construction $(E/F, \theta) \mapsto \pi'(E/F, \theta)$ which parametrizes $\mathcal{A}(G'_1) = \mathcal{A}(D^\times)$ via isomorphism classes of admissible pairs of degree $[E : F]$ dividing d.

In the case where E/F is totally ramified of degree n, there is an unramified quadratic character η of $E^\times$ such that $\pi(E/F, \theta)$ corresponds to $\pi'(E/F, \eta\theta)$ via the Jacquet–Langlands correspondence [35, when n is prime]. A similar answer is expected in general but does not seem to be available yet.

More interesting, because the construction is much more involved, is the *totally wild* case, when n is power of p and we consider $\pi \in \mathcal{A}^o(G_n)$ such that there is no unramified character χ of $F^\times$ for which $\chi\pi$ is isomorphic to π. In that case [15] π is constructed from a so-called "*simple*" pair (β, θ) where β is an element of G generating a totally ramified extension E/F of degree n; to such an element β are attached two open subgroups $J = J(\beta)$ and $H^1 = H^1(\beta)$, θ is a character of H^1 of a very specific shape, and π is induced from a representation λ of J with restriction to H^1 isotypic of type θ. There certainly exists $\beta' \in G'$ with $P(\beta) = P(\beta')$ and there is a natural procedure to deduce from θ a similar character θ' of some open compact subgroup $H^1(\beta')$ of $D^\times$ and then from λ a similar representation λ' of an open compact subgroup $J(\beta')$ which induces to $\pi' \in \mathcal{A}(D^\times)$. The main result of [11] (see [10] when $n = p$) states that, at least when p is odd, π corresponds to π' under the Jacquet–Langlands correspondence.

8. Construction and explicit Jacquet–Langlands correspondence, $1 < r < n$

Generalizing the above constructions to intermediate cases $1 < r < n$ is not easy, even when restricting to tame or totally wild situations! Moreover one wants to deal with L^2 representations and not only with cuspidal ones.

This is already visible in the so-called "*level-zero*" case, where one considers the Jacquet–Langlands correspondence $\pi \leftrightarrow \pi'$ for π corresponding to a tamely ramified representation of W_F: in $G'_r = \mathrm{GL}_r(D)$, this translates into the property that in the space of π' there is a non-zero vector fixed under the compact open subgroup $1 + P_D M_r(O_D)$ where O_D is the ring of integers of D and P_D its maximal ideal. We say that such $\pi' \in \mathcal{A}^d(G'_r)$ have *level zero*. Level zero elements of $\mathcal{A}^d(G'_r)$ or $\mathcal{A}^d(G_n)$, $n = rd$, are parametrized by admissible pairs $(E/F, \theta)$, where E/F is unramified of degree dividing n, and θ is a tamely ramified character of $E^\times$. However

it is hard to label exactly the representation corresponding to $(E/F, \theta)$, and even more difficult to get the Jacquet–Langlands correspondence explicit in that level zero case. That was recently accomplished by Grabitz, Silberger and Zink [26], [56].

The higher level case is even more difficult, and has been subject to investigations of P. Broussous and M. Grabitz [17], [18], [25] and V. Sécherre [51], [52], [53]. By analogy with [61] and [15], they construct "*simple*" *pairs* (β, θ) as above in G'_r. Sécherre then follows the procedure of [15], and constructs [51], [52] from a simple pair (β, θ) a *simple type* (J, λ) in G'_r as in § 7, and controls its intertwining, so that in particular it is known when λ gives rise by induction to cuspidal smooth irreducible representations of G'_r; in the contrary case, Sécherre computes the intertwining algebra, which in particular tells how many elements in $\mathcal{A}^d(G'_r)$ one can get from λ [53]. It remains to prove that all of $\mathcal{A}^d(G'_r)$ has been obtained: this is the problem of *exhaustion*. Grabitz has obtained some partial results in that direction [25].

This is only for the explicit construction of $\mathcal{A}^d(G'_r)$ in the higher level case: describing the Jacquet–Langlands correspondence explicitly is harder.

As a final remark, let me mention that once exhaustion is proved, Sécherre's result on the intertwining algebra will imply the conjecture of Tadić mentioned in § 4.

References

[1] Aubert, A.-M., Duality in the Grothendieck group of the category of finite length smooth representations of a p-adic reductive group. *Trans. Amer. Math. Soc.* **347** (1995), 2179–2189; Erratum ibid. *Trans. Amer. Math. Soc.* **348** (1996), 4687–4690.

[2] Badulescu, A., Orthogonalité des caractères pour GL(n) sur un corps de caractéristique non nulle. *Manuscripta Math.* **101** (2000), 49–70.

[3] Badulescu, A., Correspondance de Jacquet-Langlands pour les corps locaux de caractéristique non nulle. *Ann. Sci. École Norm. Sup.* (4) **35** (2002), 695–747.

[4] Badulescu, A., Correspondance de Jacquet-Langlands étendue à toutes les représentations. Preprint, 2005.

[5] Badulescu, A., Global Jacquet-Langlands correspondence. Preprint, 2005.

[6] Badulescu, A., Renard, D., Sur une conjecture de Tadić. *Glas. Math. Sec.* III **39** (2004), 49–54.

[7] Bernstein, J. N., P-invariant distributions on GL(N) and the classification of unitary representations of GL(N) (non-Archimedean case). In *Lie groups and representations* II, Lecture Notes in Math 1041, Springer-Verlag, Berlin 1987, 50–102.

[16] Broussous, P., Extension du formalisme de Bushnell et Kutzko au cas d'une algèbre à division. *Proc. London Math. Soc.* (3) **77** (1998), 292–326.

[17] Broussous, P., Minimal strata for GL(m, D). *J. Reine Angew. Math.* **514** (1999), 199–236.

[18] Broussous, P., Grabitz, M., Pure elements and intertwining classes of simple strata in local central simple algebras. *Comm. Algebra* **28** (2000), 5405–5442.

[8] Bushnell, C. J., Smooth representations of p-adic groups: the rôle of compact open subgroups. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 770–779.

[9] Bushnell, C. J., Henniart, G., Davenport-Hasse relations and an explicit Langlands correspondence II: twisting conjectures. *J. Théor. Nombres Bordeaux* **12** (2000), 309–347.

[10] Bushnell, C. J., Henniart, G., Correspondance de Jacquet-Langlands explicite II. Le cas de degré égal à la caractéristique résiduelle. *Manuscripta Math.* **102** (2000), 211–225.

[11] Bushnell, C. J., Henniart, G., Local tame lifting for GL_n, III. Explicit base change and Jacquet-Langlands correspondence. *J. Reine Angew. Math.* **580** (2005), 39–100.

[12] Bushnell, C. J., Henniart, G., The essentially tame local Langlands correspondence I. *J. Amer. Math. Soc.* **18** (2005), 685–710.

[13] Bushnell, C. J., Henniart, G., The essentially tame local Langlands correspondence II. *Compositio Math.* **141** (2005), 979–1011.

[14] Bushnell, C. J., Henniart, G., *The local Langlands conjecture for* GL(2). To appear.

[15] Bushnell, C. J., Kutzko, P. C., *The admissible dual of* GL(N) *via compact open subgroups*. Ann. of Math. Stud. 129, Princeton University Press, Princeton, NJ, 1993.

[19] Carayol, H., Représentations cuspidales du groupe linéaire. *Ann. Sci. École Norm. Sup.* (4) **17** (1984), 191–226.

[20] Deligne, P., Les constantes des équations fonctionnelles des fonctions L. In *Modular functions of one variable* II, Lecture Notes in Math. 349, Springer-Verlag, Berlin (1973), 501–597.

[21] Deligne, P., Kazhdan, D., Vignéras, M.-F., Représentations des algèbres centrales simples p-adiques. In *Représentations des groupes réductifs sur un corps local*, Hermann, Paris 1984, 33–117.

[22] Flath, D., A comparison of the automorphic representations of GL(3) and its twisted forms. *Pacific J. Math.* **97** (1981), 373–402.

[23] Gelbart, S., Jacquet, H., Forms of GL(2) from the analytic point of view. In *Automorphic forms, representations and L-functions*, Part I, Proc. Sympos. Pure Math. 33, Amer. Math. Soc., Providence, RI, 1979, 213–251.

[24] Gérardin, P., Li, W., Fourier transforms of representations of quaternions. *J. Reine Angew. Math.* **359** (1985), 125–173.

[25] Grabitz, M., Simple characters for principal orders and their matching I. Peprint, MPI Bonn 99-117; II, preprint, MPI Bonn 03-56; III, preprint, 2005.

[26] Grabitz, M., Silberger, A., Zink, E. W., Level zero types and Hecke algebras for local central simple algebras. *J. Number Theory* **91** (2001), 92–125.

[27] Harish-Chandra, *Admissible invariant distributions on reductive p-adic groups*. Univ. Lecture Ser. 16, Amer. Math. Soc., Providence, RI, 1999.

[28] Harish-Chandra, A submersion principle and its applications. In *Geometry and analysis*, Indian Acad. Sci., Bangalore 1980, 95–102.

[29] Harris, M., The local Langlands conjecture for GL_n over a p-adic field, $n < p$. *Invent. Math.* **134** (1998), 177–210.

[30] Harris, M., On the local Langlands correspondence. *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 583–597.

[31] Harris, M., The local Langlands correspondence. In *Formes Automorphes* I, Astérisque **298** (2005), 17–145.

[32] Harris, M., Taylor, R. *The geometry and cohomology of some simple Shimura varieties*. Ann. of Math. Stud. 151, Princeton University Press, Princeton, NJ, 2001.

[33] Henniart, G., *La conjecture de Langlands locale pour GL*(3). *Mém. Soc. Math. France* (*NS*) **11/12** (1984), 186 pp.

[34] Henniart, G., La conjecture de Langlands locale numérique pour GL(n). *Ann. Sci. École Norm. Sup* (4) **21** (1988), 497–544.

[35] Henniart, G., Correspondance de Jacquet-Langlands explicite I: le cas modéré de degré premier. In *Séminaire de théorie des nombres* (Paris, 1990–1991), Progr. Math. 108, Birkhäuser, Basel 1993, 85–114.

[36] Henniart, G., Une preuve simple des conjectures de Langlands pour GL(n) sur un corps p-adique. *Invent. Math.* **139** (2000), 439–455.

[37] Henniart, G., Correspondance de Langlands et fonctions L des carrés extérieur et symétrique. Preprint, IHES/M/03/20, 2003.

[38] Howe, R., Tamely ramified supercuspidal representations of GL(n). *Pacific J. Math.* **73** (1977), 437–460.

[39] Jacquet, H., Langlands, R. P., *Automorphic forms on* GL(2). Lecture Notes in Math. 114, Springer-Verlag, Berlin 1970.

[40] Jacquet, H., Piatetski-Shapiro, I. I., Shalika, J., Rankin-Selberg convolutions. *Amer. J. Math.* **105** (1983), 367–464.

[41] Kutzko, P. C., The Langlands conjecture for GL(2) of a local field. *Ann. of Math.* (2) **112** (1980), 381–412.

[42] Kutzko, P. C., The exceptional representations of GL(2). *Compositio Math.* **51** (1984), 3–14.

[43] Lafforgue, L., Chtoucas de Drinfeld et correspondance de Langlands. *Invent. Math.* **147** (2002), 1–24.

[44] Lafforgue, L., Chtoucas de Drinfeld, formule des traces d'Arthur-Selberg et correspondance de Langlands. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. I, Higher Ed. Press, Beijing 2002, 383–400.

[45] Laumon, G., Rapoport, M., Stuhler, U., $\mathcal{D}$-elliptic sheaves and the Langlands correspondence. *Invent. Math.* **113** (1993), 217–338.

[46] Lemaire, B., Intégrabilité locale des caractères-distributions de GL(N, F), où F est un corps local non archimédien de caractéristique quelconque. *Compositio. Math.* **100** (1996), 41–75.

[47] Lemaire, B., Intégrabilité locale des caractères tordus de $\mathrm{GL}_r(D)$. *J. Reine Angew. Math.* **566** (2004), 1–39.

[48] Moy, A., Local constants and the tame Langlands correspondence. *Amer. J. Math.* **108** (1986), 863–930.

[49] Reimann, H., Representations of tamely ramified p-adic division and matrix algebras. *J. Number Theory* **38** (1991), 58–105.

[50] Rogawski, J., Representations of GL(n) and division algebras over a local field. *Duke Math. J.* **50** (1983), 161–196.

[51] Sécherre, V., Représentations lisses de GL(m, D) I. Caractères simples. *Bull. Soc. Math. France* **132** (2004), 327–396.

[52] Sécherre, V., Représentations lisses de GL(m, D) II: β-extensions. *Compositio Math.*, to appear.

[53] Sécherre, V., Représentations lisses de GL(m, D) III: types simples. *Ann. Sci. École Norm. Sup. Paris*, to appear.

[54] Shahidi, F., Fourier transforms of intertwining operators and Plancherel measures for GL(n). *Amer. J. Math.* **106** (1984), 67–111.

[55] Shahidi, F., A proof of Langlands' conjecture on Plancherel measure, complementary series for p-adic groups. *Ann. of Math.* **132** (1990), 273–330.

[56] Silberger, A., Zink, E.-W., Weak explicit matching for level zero discrete series of unit groups of p-adic, simple algebras. *Canad. J. Math.* **55** (2003), 353–378 (and more recent preprints).

[57] Tadić, M., Classification of unitary representations in irreducible representations of general linear group (non-Archimedean case). *Ann. Sci. École Norm. Sup.* (4) **19** (1986), 335–382.

[58] Tadić, M., Induced representations of GL(n, A) for p-adic division algebras A. *J. Reine Angew. Math.* **405** (1990), 48–77.

[59] Tate, J., Fourier Analysis in Number Fields and Hecke's Zeta functions. In *Algebraic Number Theory* (ed. by J. W. S. Cassels and A. Fröhlich), Academic Press, London 1967, 305–347.

[60] Weil, A., *Basic Number Theory*. Appendix II, Grundlehren Math. Wiss. 144, Springer-Verlag, Berlin 1974.

[61] Zelevinsky, A., Induced representations of reductive p-adic groups II. On irreducible representations of GL(n). *Ann. Sci. École Norm. Sup.* (4) **13** (1980), 165–210.

[62] Zink, E.-W., Representation theory of local division algebras. *J. Reine Angew. Math.* **428** (1992), 1–44.

Université de Paris-Sud, Mathématiques, bât. 425, UMR 8628 du CNRS, 91405 Orsay Cédex, France

An invitation to bounded cohomology

Nicolas Monod

Abstract. A selection of aspects of the theory of bounded cohomology is presented. The emphasis is on questions motivating the use of that theory as well as on some concrete issues suggested by its study. Specific topics include rigidity, bounds on characteristic classes, quasification, orbit equivalence, amenability.

Mathematics Subject Classification (2000). Primary 55N; Secondary 20F.

Keywords. Bounded cohomology.

1. Introduction

This lecture is an invitation to explore bounded cohomology. It is not an attempt to collect all recent advances in that topic, much less an ex cathedra exposition of the theory. I will try to illustrate how problems from diverse origins can be translated into the framework of bounded cohomology, and how in return this theory prompts a few concrete problems. A number of questions are suggested, ranging from teasers to the ill-defined.

The terminology *bounded cohomology* proposed by M. Gromov [67] refers to the following concrete definition. Recall that the ordinary (singular) cohomology $\mathrm{H}^*(M, \mathbf{R})$ of a manifold M can be defined by the complex of all *singular cochains* on M, which are just all real-valued functions on the set of singular simplices. The subspace of *bounded* functions yields a subcomplex and hence cohomology groups $\mathrm{H}^*_\mathrm{b}(M, \mathbf{R})$. Moreover, the inclusion map of this subcomplex determines a natural transformation $\mathrm{H}^*_\mathrm{b}(M, \mathbf{R}) \to \mathrm{H}^*(M, \mathbf{R})$ called the *comparison map*.

This definition can be imitated for groups. One of the definitions of Eilenberg–MacLane cohomology $\mathrm{H}^*(G, V)$ of a group G with coefficients in some module V is given by the complex of all G-equivariant V-valued functions on G^{n+1}. Considering only those functions that are bounded one obtains a complex

$$0 \longrightarrow \mathrm{C}_\mathrm{b}(G, V)^G \longrightarrow \mathrm{C}_\mathrm{b}(G^2, V)^G \longrightarrow \mathrm{C}_\mathrm{b}(G^3, V)^G \longrightarrow \cdots$$

whose cohomology is the *bounded cohomology* $\mathrm{H}^*_\mathrm{b}(G, V)$ and comes again with a *comparison map* $\mathrm{H}^*_\mathrm{b}(G, V) \to \mathrm{H}^*(G, V)$. One needs here to make sense of boundedness for V-valued maps; for instance, V could be a Banach space with isometric G-representation. We will usually assume that this *coefficient module* is dual (e.g.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$V = \mathbf{R}$ or a unitary representation on Hilbert space). More generally, when G is a locally compact group we (ab)use the same notations H^*, H^*_b for the *continuous* (bounded) cohomology; for present purposes, it is simply defined by requiring that all cochains be continuous[1].

I will mainly concentrate on the group case; a theorem of M. Gromov of fundamental importance states that the morphism $\mathrm{H}^*_\mathrm{b}(\pi_1(M), \mathbf{R}) \to \mathrm{H}^*_\mathrm{b}(M, \mathbf{R})$ induced by the classifying map is an isomorphism (Corollary p. 40 in [67]; see also [11]).

The above definitions, especially in the group case, might seem artificial. The next section is an attempt to challenge this perception. Nevertheless, I should point out that *there is not a single countable group G for which $\mathrm{H}^*_\mathrm{b}(G, \mathbf{R})$ is known, unless it is known to vanish in all degrees.* For instance, here is what is known for the free group $\mathbf{F}_2$ and trivial coefficients: $\mathrm{H}^1_\mathrm{b}(\mathbf{F}_2, \mathbf{R})$ vanishes, $\mathrm{H}^2_\mathrm{b}(\mathbf{F}_2, \mathbf{R})$ and $\mathrm{H}^3_\mathrm{b}(\mathbf{F}_2, \mathbf{R})$ are infinite-dimensional [11, §3], [66], [139].

However, in the case of *connected* groups, all known results seem to indicate that bounded cohomology with trivial coefficients is much better behaved than for discrete groups. The connected case can be reduced to semi-simple Lie groups, prompting the following question.

Problem A. Let G be a connected semisimple Lie group with finite centre. Is the comparison map

$$\mathrm{H}^*_\mathrm{b}(G, \mathbf{R}) \longrightarrow \mathrm{H}^*(G, \mathbf{R})$$

an isomorphism?

At this time, it seems that the question of injectivity and surjectivity of this comparison map are two quite different issues; existing proofs are of a very different nature. Perhaps this will change with a better understanding of the bounded cohomology of Lie groups. But for the time being, I would like to single out the subquestion below because it is the aspect most relevant for several questions discussed in this text.

Problem A′. Is the comparison map of Problem A surjective?

A related conjecture was proposed by J. Dupont [43]; compare also the introduction of [14] and 9.3.8, 9.3.9 in [112]. One can also ask the same questions more generally for products of semisimple algebraic groups over local fields. For more general coefficient modules, even if only irreducible unitary representations are considered, both injectivity and surjectivity fail [25].

2. Three ways to stumble upon H^*_b

I will now sketch briefly three circumstances leading naturally to study bounded cohomology: (1) group algebras; (2) bounds and refinements of classical invariants; (3) quasification.

[1] I emphasise that when G is not discrete, this is not the same as the cohomology of the classifying space BG.

2.1. Cohomology of group algebras. The usual Eilenberg–MacLane cohomology of a discrete group Γ can be seen as a particular case of the cohomology of algebras; namely, it is isomorphic to the cohomology of the group algebra $A = \mathbf{R}\Gamma$.[2] Here A is the free $\mathbf{R}$-vector space on Γ endowed with the convolution product. In other words, it is the universal object obtained by forcing upon the group Γ an $\mathbf{R}$-linear structure extending the multiplication in Γ. Recall that the cohomology of A is defined by Ext-functors, which means that one considers complexes of linear morphisms

$$A \otimes \cdots \otimes A \longrightarrow \mathbf{R}$$

to $\mathbf{R}$ or more generally to suitable coefficient modules.

Now, this free vector space A brings along its "free norm", that it, the ℓ^1-norm; the latter extends as the *projective tensor norm* on the above tensor products [72, I§1.1]. Why not, then, take this topology into account and compute the cohomology of the *topological algebra* A? This amounts to considering complexes of *continuous* linear morphisms

$$A \otimes_\pi \cdots \otimes_\pi A \longrightarrow \mathbf{R},$$

where $\otimes_\pi$ is the notation indicating the choice of the projective tensor norm. The point is that *this cohomology is nothing else than the bounded cohomology of* Γ, as follows readily from the properties of $\otimes_\pi$. Once we deal with this *continuous cohomology* of A, we may of course replace A by its completion $\ell^1(\Gamma)$ without affecting the outcome. In conclusion, *the bounded cohomology of groups is a particular case of the cohomology of Banach algebras* as exposed in B. Johnson's 1972 memoir [89] (see also A.Ya. Helemskiĭ [76]). I would like to point out that the group algebra case (that is, bounded cohomology) was indeed prominent in B. Johnson's memoir. One can find therein several aspects that became intensively studied later, such as quasimorphisms, amenability and the problem of the existence of outer derivations. Nevertheless, it is M. Gromov's paper (which also refers to ideas of W. Thurston) that gave all its impetus to the theory.

Here are two questions to conclude this outline. First, we point out that for C *-algebras and specifically von Neumann algebras, other cohomologies have been studied, such as *completely bounded cohomology*, where the notion of boundedness is quite different from our setting; see Christensen–Effros–Sinclair [33]. Consider the von Neumann algebra $L(\Gamma)$ associated to a countable group Γ, which is the completion of A for a much weaker topology than above. Then, there is not anymore a straightforward correlation between Γ-modules and $L(\Gamma)$-bimodules as there was for $\ell^1(\Gamma)$. However, there is a precise analogy instead: the theory of *correspondences* (see A. Connes [36] and S. Popa [130]). Whilst this will not provide a precise dictionary from the Eilenberg–MacLane cohomology of Γ to some cohomology of $L(\Gamma)$, it still

[2] For exact statements, one needs to give more details on how to handle the coefficients involved; for instance, A is suitable for coefficients that are real vector spaces. Moreover, Hochschild cohomology of algebras involves A-bimodules as coefficients, whilst a priori the cohomology of Γ is defined for modules. Not taking this into account would identify the Hochschild cohomology with a sum of Eilenberg–MacLane cohomology of centralisers of representatives of the conjugacy classes of Γ. We will omit all details here.

raises the interesting possibility to define for $L(\Gamma)$ an analogue of the L^2-cohomology of Γ in the sense of M. Atiyah [4], Cheeger–Gromov [31] and W. Lück [101]. Such an analogue has been proposed by A. Connes and D. Shlyakhtenko in [38].

Here is how this programme connects to the topic of our discussion. For present purposes, let us think of the L^2-cohomology of Γ as the Eilenberg–MacLane cohomology $\mathrm{H}^*(\Gamma, \ell^2(\Gamma))$, where $\ell^2(\Gamma)$ is endowed with the regular representation[3]. The *bounded* cohomology $\mathrm{H}^*_\mathrm{b}(\Gamma, \ell^2(\Gamma))$ has also proved very useful in degree two (see Sections 4.1 and 4.2), even though one cannot *measure* its size as is done by means of the von Neumann dimension in the case of $\mathrm{H}^*(\Gamma, \ell^2(\Gamma))$.

Problem B. Perform a construction analogous to Connes–Shlyakhtenko [38] but corresponding to $\mathrm{H}^*_\mathrm{b}(\Gamma, \ell^2(\Gamma))$ instead of $\mathrm{H}^*(\Gamma, \ell^2(\Gamma))$. Provide non-triviality results, at least in degree two.

The second question is more indeterminate:

Problem C. Consider the cyclic cohomology of locally convex algebras $\mathcal{A}$ as in §II.5 of A. Connes' [37]. What can be said for $\mathcal{A} = \ell^1(\Gamma)$?

The cyclic homology of the group algebra $A = \mathbf{R}\Gamma$ (devoid of any topology) has been studied notably by D. Burghelea [29]. For an example with a locally convex completion of A, smaller than $\ell^1(\Gamma)$ and closer to the spirit of [37], see R. Ji [87].

2.2. Refinements of ordinary cohomology and numerical bounds. Certain classical cohomology classes are given by explicit cocyles that happen to be bounded. This is of particular interest for characteristic classes, since explicit bounds on characteristic numbers of flat bundles carry important geometric information, such as in the Milnor–Wood inequality; we refer to J. Dupont [43] for more on such bounds.

Here is an example of this situation: The *Euler number* of a flat oriented n-vector bundle over a compact manifold M measures the obstruction to finding a non-vanishing section. It can be computed by triangulating M, choosing generic "affine" sections over the resulting simplices and then adding up obstruction signs ± 1 for each simplex – see D. Sullivan [141] and J. Smillie [138]; compare also [7, F.4]. The resulting number is clearly bounded in terms of the number of simplices, though this bound is mysterious. One obtains a nice conceptual bound by observing that the Euler class of $\mathrm{GL}^+_n(\mathbf{R})$ has an explicit cocycle representative which is bounded, e.g. as in Ivanov–Turaev [86] (GL^+_n refers to the group of matrices with positive determinant).

Problem D. Let $\pi = (B \to M)$ be a flat oriented n-vector bundle over a compact manifold M. Given additional structure on π, define a natural class $\mathcal{E}_\mathrm{b}(\pi)$ in $\mathrm{H}^n_\mathrm{b}(M, \mathbf{R})$ whose image in $\mathrm{H}^n(M, \mathbf{R})$ is the Euler class $\mathcal{E}(\pi)$ of π.

[3] This is correct under the finiteness assumption F_n, where n is the degree of the cohomology considered. For the general case, one should follow W. Lück [101].

In particular, given a compact orientable n-manifold M with additional structure, find a natural definition of $\mathcal{E}_{\mathrm{b}}(M) \stackrel{\text{def}}{=} \mathcal{E}_{\mathrm{b}}(TM \to M)$.

I should certainly be a bit more specific here, since after all one way to see the Milnor–Wood inequality is to consider an explicit cocycle for the Euler class $\mathcal{E}$ that witnesses its boundedness. However, this is quite different from having a *natural invariant* in $\mathrm{H}^n_{\mathrm{b}}(M, \mathbf{R})$, already because the map $\mathrm{H}^n_{\mathrm{b}}(M, \mathbf{R}) \to \mathrm{H}^n(M, \mathbf{R})$ can be far from injective. The fact that additional structure on M should be required to rigidify the situation is suggested by the fact that $\mathrm{H}^n_{\mathrm{b}}(M, \mathbf{R})$ is canonically isometrically isomorphic to the bounded cohomology of the fundamental group $\Gamma = \pi_1(M)$ [67, p. 40]. By contrast, the Euler class *as class of* $B\mathrm{GL}^+_n(\mathbf{R})$ is unbounded (compare also footnote 1); thus the idea would be to transit via the *group cohomology*:

Consider for instance the setting of compact orientable manifolds supporting an affine structure. In that case the Euler class $\mathcal{E}(M)$ will come from the structure group $\mathrm{GL}^+_n(\mathbf{R})$ by pull-back through the holonomy representation of $\pi_1(M)$ via $\mathrm{H}^n(\pi_1(M), \mathbf{R})$ and the classifying map. In that case one would have a completely canonical choice for $\mathcal{E}_{\mathrm{b}}(M)$ if, as suggested by Problem A, one proves

$$\mathrm{H}^*_{\mathrm{b}}(\mathrm{PSL}_n(\mathbf{R}), \mathbf{R}) \cong \mathrm{H}^*(\mathrm{PSL}_n(\mathbf{R}), \mathbf{R}).$$

Indeed, an easy argument shows that this would imply

$$\mathrm{H}^n_{\mathrm{b}}(\mathrm{GL}^+_n(\mathbf{R}), \mathbf{R}) \cong \mathrm{H}^n(\mathrm{GL}^+_n(\mathbf{R}), \mathbf{R}), \quad n \neq 1.$$

A motivation for Problem D is the following.

Problem D′. Use a natural definition of $\mathcal{E}_{\mathrm{b}}(M)$ to prove that, for any compact orientable manifold M supporting an affine structure, $\mathcal{E}_{\mathrm{b}}(M)$ vanishes.

This would settle the Chern–Sullivan problem of the vanishing of the Euler-Poincaré number of such manifolds M.

The example of the boundedness of the Euler class can be considerably generalised: M. Gromov proves in [67] that all primary[4] characteristic classes are bounded, at least when viewed as classes of the structure group made discrete. A different proof avoiding the use of H. Hironaka's resolution of singularities was provided by M. Bucher-Karlsson in her thesis [14]; it is also shown in [14, p. 60] how it follows that these primary characteristic classes are bounded already when viewed as classes of the topological group G. In other words, they lie in the image of the comparison map $\mathrm{H}^*_{\mathrm{b}}(G, \mathbf{R}) \to \mathrm{H}^*(G, \mathbf{R})$.

Problem E. Prove the same statement for secondary characteristic classes[5].

As pointed out in [14], this would then solve Problem A′. Indeed, J. Dupont and F. Kamber proved in [44] that, as an algebra, $\mathrm{H}^*(G, \mathbf{R})$ is generated by primary and

[4] More precisely, the characteristic classes of flat G-bundles, where G is an algebraic subgroup of $\mathrm{GL}_n(\mathbf{R})$.

[5] See Cheeger–Simons [32] and M. Bucher-Karlsson [14] for the definition of secondary classes.

secondary classes when G is a connected semisimple Lie group with finite centre. (The product is easily seen to preserve boundedness of cohomology classes.)

Once a natural bounded representative has been identified for a classical cohomological invariant, this opens the door to a refined invariant: Indeed, the *bounded class* of that bounded cocycle contains a priori much more information than the class one started with. This is because it is easier for cocycles to be cohomologous than to be "boundedly cohomologous", that is, equivalent modulo *bounded coboundaries*.

A beautiful illustration of this phenomenon has been given by É. Ghys in [63] and goes as follows. Recall that if a group Γ acts by orientation-preserving homeomorphisms on the circle, it inherits an Euler class in $\mathrm{H}^2(\Gamma, \mathbf{Z})$. This class has a canonical representative taking only values $\{0, 1\}$, thus determining a bounded class. Whilst the original class determines only the obstruction to lifting the Γ-action to an action on the line, É. Ghys proves that the bounded class completely characterises the action up to semi-conjugacy.

Another important benefit of identifying an ordinary cohomology class as being bounded is the *Gromov seminorm*. Since *homology* classes are given by cycles, i.e. formal *finite* linear combinations of simplices (singular or otherwise), there is a numerical invariant attached to every homological class, the Gromov seminorm, which is by definition the infimum of the total mass of all linear combinations representing that class. This is a method of assigning a *numerical invariant* to cohomological invariants. For instance, this number is computed for the Kähler class by Domic–Toledo [42] and Clerc–Ørsted [34].

In particular, M. Gromov [67] defines the *simplicial volume* of a closed (connected, orientable) manifold M to be the seminorm of its fundamental class (for the relative case, see also [99]). This invariant, besides its own interest, provides bounds for the minimal volume of M over all (suitably normalised) Riemannian structures, as explained by M. Gromov in [67]. The relevance of bounded cohomology is that any upper bound on a *cocycle* with non-trivial homological pairing on the given cycle provides a non-trivial lower bound on the Gromov seminorm of the cycle.

W. Thurston and M. Gromov (see the 1978 notes [142] and [67]) have shown how to use negative curvature to prove the boundedness of fundamental classes – equivalently, the positivity of the simplicial volume. The case of symmetric spaces has been solved only quite recently, with J.-F. Lafont and B. Schmidt proving in [100] that *the simplicial volume of any closed locally symmetric space of non-compact type is positive*. (A crucial ingredient of their proof is the work of C. Connell and B. Farb [35].) Actually, the case of the symmetric space of $\mathrm{SL}_3(\mathbf{R})$ is not covered by [100]; it was claimed in [134], but the proof therein is incomplete. However a proof is provided by M. Bucher-Karlsson in [13]. M. Gromov has conjectured more generally that closed manifolds with non-positive curvature and negative Ricci curvature have positive simplicial volume.

2.3. Quasification. The word *quasification* is meant to refer to the process whereby a geometric or algebraic notion is modified to an approximate variant; typically, a

defining equality or inequality is relaxed by imposing that it hold up to some constants only. Here are a few examples of this overly vague principle.

(i) *Quasi-isometries.* Whereas a map $f\colon X \to Y$ between metric spaces X, Y is called isometric if $d_Y(f(x), f(x'))$ equals $d_X(x, x')$ for all $x, x' \in X$, it is said to be *quasi-isometric* if there is some constant $C \neq 0$ such that

$$C^{-1}d_X(x, x') - C \;\leq\; d_Y(f(x), f(x')) \;\leq\; Cd_X(x, x') + C \quad \text{for all } x, x' \in X.$$

A *quasi-isometry* is a quasi-isometric map whose image has finite codiameter. Here are two important motivations for this notion: (1) Geometric rigidity questions such as Mostow's strong rigidity [119] lead to consider quasi-isometries of symmetric spaces arising from a homotopy equivalence between two of its compact quotients; (2) Geometric group theory has had considerable success, following M. Gromov, in viewing finitely generated groups as metric spaces and studying their geometry [68], [71]; however, this point of view is well-defined only up to quasi-isometry.

As a matter of terminology, a *rough isometry* shall be a quasi-isometry where additive constants only are allowed:

$$d_X(x, x') - C \;\leq\; d_Y(f(x), f(x')) \;\leq\; d_X(x, x') + C.$$

(ii) *Gromov-hyperbolicity.* In the context of point (2) in (i) above, one calls a geodesic metric space *Gromov-hyperbolic* if any finite configuration of points is roughly isometric to a configuration in a tree; the additive constant is allowed to depend on the space and the number of points only. (For comparison with more usual definitions, see 2 §2, Theorem 12 (ii) in [64].) This notion is a quasi-isometry invariant even though it is a priori defined by rough isometries.

The theory of those finitely generated groups that are Gromov-hyperbolic as metric spaces is one of the major contribution to modern group theory; we refer to M. Gromov [69].

(iii) *Hyers–Ulam stability.* D. Hyers observed in [80] that whenever a mapping f between Banach spaces satisfies the additive equation upon some constant δ, then f is at finite distance (at most δ) of a truly additive mapping. A subsequent joint paper with S. Ulam [81] proposed the more delicate question of the stability of the equation defining isometries; more precisely: *If f is a rough isometry, is it close to an isometry?* After many partial results spanning half a century (starting with the Hilbertian case in [81]), the general case was solved affirmatively by P. Gruber [73] and J. Gevirtz [62] and the sharp constant provided by M. Omladič and P. Šemrl [122] in 1995.

The stability question was broadened to other contexts by Hyers–Ulam [82], [83] and a great many other authors. Within the context of isometries, three examples are: quaternionic hyperbolic spaces (P. Pansu [125]), higher rank symmetric spaces and buildings (Eskin–Farb [47] and Kleiner–Leeb [97]), hyperbolic buildings (Bourdon–Pajot [10]).

The connection between quasification and bounded cohomology appears when one considers the stability of *cocycles*. Indeed, suppose that f is "almost an n-cocycle"

for a group G in the Hyers–Ulam sense, say for a Banach G-module V. Specifically, in the model of the *inhomogeneous* bar-resolution, f is a map $G^n \to V$ with the property that

$$g_1 f(g_2, \dots, g_{n+1}) + \sum_{j=1}^{n} (-1)^j f(g_1, \dots, g_j g_{j+1}, \dots g_{n+1}) + (-1)^{n+1} f(g_1, \dots, g_n)$$

is bounded independently of $g_1, \dots g_{n+1} \in G$. The map $\delta f \colon G^{n+1} \to V$ defined by the above expression is therefore a bounded $(n+1)$-cocycle; indeed it is certainly a cocycle since it is defined as a coboundary. The latter observation means that δf represents a trivial class in usual cohomology $\mathrm{H}^{n+1}(G, V)$. But is it trivial as bounded cohomology class in $\mathrm{H}_{\mathrm{b}}^{n+1}(G, V)$? The definition of bounded cohomology gives us the answer: this class is trivial if and only if f is at finite distance of an actual n-cocycle $G^n \to V$. In conclusion, the Hyers–Ulam stability problem for $\mathrm{H}^n(G, V)$ is exactly captured by the kernel of the comparison map

$$\mathrm{EH}_{\mathrm{b}}^{n+1}(G, V) \stackrel{\text{def}}{=} \mathrm{Ker}\big(\mathrm{H}_{\mathrm{b}}^{n+1}(G, V) \longrightarrow \mathrm{H}^{n+1}(G, V)\big)$$

in one degree higher: $\mathrm{EH}_{\mathrm{b}}^{n+1}$ describes "n-quasicocycles". This could be made formal by introducing suitably complexes of quasicocycles and defining the corresponding cohomology groups $\mathrm{H}^*_{\mathrm{quasi}}(G, V)$. It is then straightforward to verify that one has an infinite exact sequence

$$\begin{aligned} \cdots \to \mathrm{H}^{n-1}_{\mathrm{quasi}}(G, V) \to \mathrm{H}^n_{\mathrm{b}}(G, V) \to \mathrm{H}^n(G, V) \\ \to \mathrm{H}^n_{\mathrm{quasi}}(G, V) \to \mathrm{H}^{n+1}_{\mathrm{b}}(G, V) \to \cdots \end{aligned}$$

Consider the simplest case, namely $n = 1$ and $V = \mathbf{R}$:

Definition 2.1. A *quasimorphism* is a map $f \colon G \to \mathbf{R}$ such that

$$\sup_{g,h \in G} |f(g) - f(gh) + f(h)| < \infty.$$

A quasimorphism is *non-trivial* if it is not a bounded perturbation of a homomorphism, or equivalently if it determines a non-zero class in $\mathrm{H}^2_{\mathrm{b}}(G, \mathbf{R})$.

B. Johnson proves already in [89, 2.8] that the free group $\mathbf{F}_2$ admits a non-trivial quasimorphism. This was considerably generalised and it is now known that $\mathrm{EH}^2_{\mathrm{b}}(G, \mathbf{R})$ is infinite-dimensional for (non-elementary) free groups, surface groups, Gromov-hyperbolic groups, free products (R. Brooks [11], Brooks-Series [12], Y. Mitsumatsu [111], Barge–Ghys [5], Epstein–Fujiwara [46], K. Fujiwara [49], [50], R. Grigorchuk [66]); generalising all the previous cases, for all groups acting on a Gromov-hyperbolic metric space in a *weakly proper* way (Bestvina–Fujiwara [8]; see also U. Hamenstädt [74]). Moreover, J. Manning [102, 4.29] shows that if there is

any quasimorphism that is what he calls *bushy*, then, already $H_b^2(G, \mathbf{R})$ is infinite-dimensional. (Contrary to what has sometimes been suggested, $EH_b^2(G, \mathbf{R})$ may however be of finite non-zero dimension [115]; thus not every quasimorphism is bushy.) Very interesting quasimorphism of a completely different nature have been constructed by Entov–Polterovich [45], Biran–Entov–Polterovich [9], Gambaudo–Ghys [58] and P. Py [131]. There, quasification is the additional freedom that allows to extend the Calabi homomorphism as a quasimorphism to larger groups that do not admit any non-zero homomorphism.

One checks that amenable groups do not have non-trivial quasimorphisms. In a completely opposed direction, it was proved in [23], [24] that irreducible lattices in semisimple Lie groups of higher rank have no non-trivial quasimorphisms. Interestingly, this property is not quasi-isometry invariant [23, 1.7].

Increasing the generality, let us consider unitary representations V. Since $H^1(G, V)$ classifies affine isometric actions on the Hilbert space V, it follows that $EH_b^2(G, V)$ contains information about *rough G-actions* on V, namely maps ϱ from G to the affine isometry group of V such that

$$\sup_{g,h\in G} \sup_{v\in V} \| \varrho(g)(\varrho(h)v) - \varrho(gh)v \| < \infty$$

(this forces the linear part of ϱ to be an actual representation).The results of [23], [24] show that any such rough action of a higher rank lattice has bounded "orbits". Using Hyers–Ulam stability [81], we deduce the following corollary for higher rank lattices: *Every action by rough isometries (of a given constant) on a Hilbert space has bounded orbits.*

Another natural problem is to consider *ε-representations* (or *near representations*) of the group G on a Hilbert space V, that is, maps $\pi : G \to U(V)$ to the unitary group U such that

$$\sup_{g,h\in G} \| \pi(g)\pi(h) - \pi(gh) \|_{\mathrm{op}} < \varepsilon,$$

wherein the norm is now the operator norm. The corresponding stability question is now: How close is π to an actual unitary representation? (In operator norm.)

When G is finite or more generally compact (with appropriate continuity addenda), la Harpe–Karoubi proved that for every $\delta > 0$ there is $\varepsilon > 0$ such that every ε-representation of Γ is δ-close to a unitary representation [40]. This was then established for amenable groups by D. Kazhdan in [95] using an ingenious notion of ε-cocycles. This device is however not obviously related to bounded cohomology.

Problem F. Can one reformulate in terms of bounded cohomology the problem of the stability of unitary representations?

D. Kazhdan also gives an example showing that for surface groups the phenomenon of stability of unitary representations fails to hold (Theorem 2 in [95]).

Problem F′. Prove (or disprove): Let Γ be a lattice in a connected simple Lie group of real rank at least two, e.g. $\Gamma = \mathrm{SL}_3(\mathbf{Z})$. Then for every $\delta > 0$ there is $\varepsilon > 0$ such that every ε-representation of Γ is δ-close to a unitary representation.

Notice that this conjectural stability does not follow from Kazhdan's property (T); just as for bounded cohomology, a stronger rigidity property of higher rank groups needs to be used. Indeed, any group G with a non-trivial quasimorphism $f\colon G \to \mathbf{R}$ lacks the stability of unitary representations: Consider for $\eta \in \mathbf{R}$ the map $\pi : G \to \mathrm{U}(\mathbf{C})$ for which $\pi(g)$ is the multiplication by $e^{i\eta f(g)}$. This is an ε-representations when η is small enough, but will not be quite close to a representation. Now recall that any non-elementary hyperbolic group admits non-trivial quasimorphisms and that there are many hyperbolic groups with property (T).

3. The rôle of amenability

The relevance of amenability to bounded cohomology has been patent ever since B. Johnson's memoir [89], where it is shown that a locally compact group G is amenable if and only if $\mathrm{H}^n_{\mathrm{b}}(G, V)$ vanishes for all $n > 0$ and all dual Banach modules V (compare also G. Noskov [120]). Since $\mathrm{H}^n_{\mathrm{b}}(G, V)$ appears in [89] as the Banach algebra cohomology of the group algebra of G, B. Johnson uses this characterisation to *define* the amenability of general Banach algebras. This suggests to consider the "bounded-cohomology dimension" of a group (or Banach algebra); more precisely:

Definition 3.1. (i) Let $\dim^{\sharp}_{\mathrm{b}}(G) \in \mathbf{N} \cup \{\infty\}$ denote the smallest integer such that $\mathrm{H}^n_{\mathrm{b}}(G, V)$ vanishes for all $n > \dim^{\sharp}_{\mathrm{b}}(G)$ and all dual Banach modules V.

(ii) Let $\dim_{\mathrm{b}}(G) \in \mathbf{N} \cup \{\infty\}$ denote the smallest integer such that $\mathrm{H}^n_{\mathrm{b}}(G, V)$ vanishes for all $n > \dim_{\mathrm{b}}(G)$ and *all* Banach modules V.

Thus G is amenable if and only if $\dim^{\sharp}_{\mathrm{b}}(G) = 0$. There is a priori a hierarchy of increasingly weak generalisations of amenability given by $\dim^{\sharp}_{\mathrm{b}}(G) = n$, $n \in \mathbf{N}$ (compare [89], §10.10; standard homological techniques reduce the property $\dim^{\sharp}_{\mathrm{b}}(G) \leq n$ to showing vanishing in degree $n+1$ only). It is not clear whether this hierarchy is really non-trivial; we refer to Section 5.4, where it is shown for instance that $\dim^{\sharp}_{\mathrm{b}} \neq 1, 2$. The dimension $\dim_{\mathrm{b}}(G)$ seems more mysterious; see [88] for some results.

3.1. Amenable actions. Just as *proper* G-spaces are relevant to compute the usual cohomology of a group G, there is a notion of *amenable* G-spaces relevant for bounded cohomology. Recall that properness is reflected in the possibility to perform finite (or compact) averaging, at least when the coefficient modules are topological vector spaces. The notion of averaging is the naïve one when the proper G-space is a homogeneous space G/K with finite or compact isotropy $K < G$, and can be carried out e.g. using Bruhat functions in the more general case.

The idea now is that *bounded* cocycles should allow averaging under more general circumstances: after all, the definition of amenable groups is that they permit equivariant averaging of bounded functions. The analogue of properness should accordingly generalise homogeneous G-spaces G/K with *amenable* isotropy group $K < G$. This is precisely the notion introduced by R. Zimmer [145], [146]: A Lebesgue space (S, ν) with non-singular G-action is called *amenable* if (i) the stabiliser of ν-almost every point is an amenable subgroup of G, (ii) the equivalence relation on S induced by the action is amenable. (This, however, is not Zimmer's original formulation.)

An important feature of this approach is that one has to work within the measurable category, because the general averaging process arising from amenability does not preserve continuity.

As suggested by the analogy with properness, one has the following result: *The G-space S is amenable if and only if the G-module $L^\infty(S)$ is relatively injective* in a sense suitably adapted to bounded cohomology [24], [112]. It then follows from functorial machinery that the bounded cohomology $\mathrm{H}_\mathrm{b}^*(G, V)$ of a locally compact group G in a coefficient module V is canonically realised by the complex

$$0 \longrightarrow L^\infty(S, V)^G \longrightarrow L^\infty(S^2, V)^G \longrightarrow L^\infty(S^3, V)^G \longrightarrow \cdots .$$

Such a statement does require a functorial theory for the bounded cohomology of groups. Even though H_b^2 lacks the basic properties of cohomological functors, such a machinery has been developed; see R. Brooks [11], N. Ivanov [85], G. Noskov [120], [121] for discrete groups and [24], [112] for locally compact groups and for the connection with amenable spaces.

All this would not be very useful without interesting examples of amenable spaces; the foremost example is provided by *Poisson boundaries* of random walks. Recall that a random walk on G is given by a probability μ on G, which for simplicity we assume *full*, that is: (i) μ is absolutely continuous with respect to Haar measures, (ii) the support of μ generates G as a semi-group. To such a random walk one associates a non-singular G-space $S = \partial_\mu G$, the *Poisson boundary*; see H. Furstenberg [53], [54], [55], Kaĭmanovich–Vershik [94], V. Kaĭmanovich [90], A. Furman [52]. R. Zimmer proved that this G-space is amenable [144], [145]. (For another proof, see [92].)

One reason why the Poisson boundary $S = \partial_\mu G$ is a useful example of amenable space is that it is much "smaller" than the only obvious amenable G-space, G itself. Specifically, for μ symmetric, the diagonal action on S^2 is ergodic, as shown by L. Garnett [60, Remark p. 301] (generalising an argument which goes back to a 1939 paper of E. Hopf [79]). In fact, it satisfies even a much stronger double ergodicity property introduced in [24], [112]: *Every G-equivariant measurable map on S^2 to every continuous separable Banach G-module is constant.* (The existence of random walks with this property was established in [24], whilst the general – and nicer – proof was later provided by V. Kaĭmanovich in [91].) If we consider the above complex, we deduce that in this situation we have a canonical identification

$$\mathrm{H}_\mathrm{b}^2(G, V) \;\cong\; \big\{\text{cocycles in } L^\infty(S^3, V)^G\big\}/\{\text{constants}\}.$$

This concrete realisation is one of the most useful facts for studying bounded cohomology, as it allows to control explicitly whether or not a cocycle represents a non-vanishing class. Not only is this crucial to prove vanishing as well as non-vanishing theorems; it is also the main ingredient to prove cohomological statements such as the splitting

$$\mathrm{H}^2_{\mathrm{b}}(G_1 \times G_2, V) \;\cong\; \mathrm{H}^2_{\mathrm{b}}(G_1, V^{G_2}) \oplus \mathrm{H}^2_{\mathrm{b}}(G_2, V^{G_1})$$

for product groups $G = G_1 \times G_2$, see [24]. I emphasise that the occurrence of the space of G_i-invariants V^{G_i} is what distinguishes this statement from a mere Künneth formula and makes it consequential for rigidity applications (just as Y. Shalom's splitting formula for usual cohomology implies rigidity statements in [136]).

3.2. Amenability degree. The remarkable properties of the Poisson boundary suggest the following notion.

Definition 3.2. Let G be a countable group (or more generally a locally compact σ-compact group). Define the *amenability degree* $a(G) \in \mathbf{N} \cup \{\infty\}$ to be the supremum of all integers n for which there is some amenable G-space (X, μ) such that the diagonal G-action on X^n has finitely many ergodic components.

For an amenable group G, the amenability degree is $a(G) = \infty$ since one can take X to be a point. On the other hand, the properties of Poisson boundaries show that $a(G) \geq 2$ for *any* G. We claim: *Every non-elementary Gromov-hyperbolic group G satisfies $a(G) = 2$.* Indeed, this holds more generally[6] for all groups G with infinite-dimensional $\mathrm{H}^2_{\mathrm{b}}(G, \mathbf{R})$ in view of the discussion in Section 3.1:

Proposition 3.3. *If G has infinite-dimensional $\mathrm{H}^n_{\mathrm{b}}(G, \mathbf{R})$, then $a(G) \leq n$.*

Problem G. What are the possible values of $a(G)$? Can one have $3 < a(G) < \infty$?

Note that $G = \mathrm{PSL}_2(\mathbf{R})$ satisfies $a(G) \geq 3$ in view of its canonical action on the projective line. I have no example ready of a countable group G with $2 < a(G) < \infty$.

Problem H. Does $a(G) = \infty$ imply that G is amenable?

The latter question has a connection to an old question[7] about R. Thompson's group F, namely: *is it is or is it ain't amenable?* A positive answer to Problem H would prove that F is non-amenable.

Proof. Recall first that F can be defined as the group of all piecewise affine homeomorphisms of the interval $[0, 1]$ with finitely many breakpoints at dyadic rationals and whose slopes are all powers of two [30]. The similar definition with $X = \mathbf{R}/\mathbf{Z}$ instead of $[0, 1]$ yields a group T whose diagonal action on X^n is ergodic for all n (indeed, its

[6] But the hyperbolic case can also be treated by a more geometric argument.

[7] Reportedly already considered by R. Thompson in the sixties, and then independently asked by R. Geoghegan in 1979.

action on dyadic rationals is oligomorphic). The stabiliser of any dyadic point of X is isomorphic to F, whilst the stabiliser of a non-dyadic point is an increasing union of groups isomorphic to F. Thus, if F is amenable, every stabiliser is amenable. On the other hand, the equivalence relation of the T-action on X can be seen to be hyperfinite in the Borel sense. It follows that the T-action on X would be amenable with respect to any quasi-invariant measure, and $a(T) = \infty$ would follow. On the other hand, it is well-known and easy to verify that T contains non-Abelian free subgroups, hence is non-amenable. (The above line of reasoning would actually imply that the T-action on $\mathbf{R}/\mathbf{Z}$ is amenable in the topological sense of [3].) □

A first step towards Problems G and H could be the following question.

Problem I. Let B be the Poisson boundary of a symmetric (full) random walk on the group G. Suppose that the diagonal G-action on B^4 is ergodic. Does it follow that G is amenable?

One can perhaps investigate this question by considering the space of *pairs of bi-infinite random paths* on G, where a bi-infinite random path refers to a random sequence

$$(\dots, x_{-2}, x_{-1}, x_0, x_1, \dots, x_n, \dots), \quad x_i \in G,\ i \in \mathbf{Z},$$

where all increments $x_n^{-1}x_{n+1}$ are i.i.d. according to the random walk (and, say, x_0 follows Haar measure class). The diagonal G-action commutes with the shift of indices and one can try to construct invariants of pairs of such paths under some assumption similar to non-recurrence of the random walk. For instance, consider the length in $\mathbf{N}$, or location in G, of the shortest segment in a Cayley graph connecting two paths; even equivariant G-valued "invariants" prevent higher ergodicity. Whilst there is a natural map from pairs of bi-infinite paths to B^4, it is not clear whether Problem I can be solved in this way.

4. Rigidity

Bounded cohomology has proved to be very useful to establish rigidity results. One can distinguish roughly three settings: obstructions, invariants, superrigidity.

(i) *Obstructions.* The idea here is simply to play off vanishing against non-vanishing; that is, to prove that for certain groups Γ, H there can be no non-trivial homomorphism $\Gamma \to H$ because (a) Γ has a vanishing property for $\mathrm{H}_{\mathrm{b}}^*$ and (b) H and certain of its subgroups have non-zero classes in $\mathrm{H}_{\mathrm{b}}^*$. Of course, the sense in which homomorphisms are non-trivial need to be precised and will affect which subgroups of H are considered. Interestingly, parts (a) and (b) are in general of a completely different nature and are proved by very different means (and often by different authors).

Example 4.1. This rather coarse strategy can be quite effective. For instance, if Γ is a higher rank lattice, it can be used to re-prove the following result of Farb–Kaĭmanovich–Masur [93], [48]: *Every representation of* Γ *into any mapping class*

group has finite image. Indeed, Γ has no non-trivial quasimorphisms [23], [24]. On the other hand, by [8] any non virtually Abelian subgroup of mapping class groups has an infinite-dimensional space of quasimorphisms. It follows that any image of Γ in a mapping class group is virtually Abelian and hence finite. (Notice that this proof does not use Margulis' normal subgroup theorem.)

(ii) *Invariants.* Sometimes one particular invariant in $\mathrm{H}_{\mathrm{b}}^*$ classifies homomorphisms. A first example appeared in Section 2.2 with É. Ghys' study of actions on the circle up to semi-conjugacy. Another is A. Iozzi's proof [84] of Matsumoto's theorem [106]. Here is a further instance due to Burger–Iozzi [18] and Burger–Iozzi–Wienhard [20]:

Example 4.2. Let X be an irreducible Hermitian symmetric space not of tube type, and let $H = \mathrm{Is}^0(X)$ be (the connected component of) its isometry group. By [23], $\mathrm{H}_{\mathrm{b}}^2(H, \mathbf{R})$ is generated by a bounded representative ω of the Kähler class of X. It is proved in [20] that for any finitely generated group Γ, the Zariski-dense representations $\pi\colon \Gamma \to H$ are classified up to conjugacy by the invariant $\pi^*\omega$ in $\mathrm{H}_{\mathrm{b}}^2(\Gamma, \mathbf{R})$. (The case $H = \mathrm{SU}(p, q)$ was previously proved in [18].)

A more refined analysis is possible for representations of surface groups by considering the *Toledo number* associated to the pull-back of the Kähler class; see the study of maximal representations presented by Burger–Iozzi–Wienhard [22] and some of its consequences [19]. Of particular importance in this context are *tight homomorphisms*, namely those representations π for which π^* preserves the norm of the Kähler class [21], [143].

(iii) *Superrigidity.* The use of richer coefficient modules, specifically of the *regular representation* $V = \ell^2(\Gamma)$, allows in some cases to encode the entire geometric situation into $\mathrm{H}_{\mathrm{b}}^2(\Gamma, V)$. This is described in the next two sections.

Non-trivial coefficients of L^∞ type have been used in [113] to establish cohomological stabilization of the general linear groups. They have also been used by Burger–Iozzi to study representations that are maximal for yet another invariant, the *generalized Toledo number*, establishing deformation rigidity for representations into $\mathrm{SU}(m, 1)$ of lattices in $\mathrm{SU}(n, 1)$ ($m \geq n \geq 2$), extending the famous result of Goldman–Millson [65] to *non-uniform* lattices [15], [16], [17].

4.1. Negative curvature made cohomological. M. Gromov suggests in [71, $7.\mathrm{E}_1$] how to turn the thin triangle property of negatively curved manifolds into a cohomological invariant (and refers to Z. Sela [135]); in his construction, Kazhdan's property (T) is used to ensure non-triviality. Building on similar ideas and using the multiple ergodicity of Poisson boundaries (Section 3.1), it is shown in [117], [110] that one has $\mathrm{H}_{\mathrm{b}}^2(\Gamma, \ell^2(\Gamma)) \neq 0$ for a large class of "negatively curved" groups Γ (see also [75]).

This result can be combined with rigidity techniques (such as Furstenberg maps) and cohomological results from [24], [112] in order to obtain superrigidity theorems. Specifically, it can be fed into the splitting formula at the end of Section 3.1 to obtain

geometric information. As a first illustration, consider the result below; here $\mathrm{H}_{\mathrm{b}}^2$ is used as a tool to control completely all existing homomorphisms, rather than just provide an obstruction when no homomorphism exists.

Let $\Gamma < G = G_1 \times G_2$ be a lattice in a product of arbitrary locally compact groups that is *irreducible* in the sense that its projection to each G_i is dense. Let $H = \mathrm{Isom}(X)$ be the isometry group of a metric space X that is negatively curved in the sense that it is either a proper CAT(-1) space or a Gromov-hyperbolic graph of bounded valency.

Theorem ([116, 1.5]). *Any non-elementary homomorphism $\Gamma \to H$ extends to a continuous homomorphism $G \to H$ (which must factor through some G_i), possibly after factoring out a compact normal subgroup of H.*

A similar statement holds more generally for cocycles in the sense of R. Zimmer [146].

(Compare [117], [110].) For previous results concerning algebraic groups, see Margulis [103], [104], Burger–Mozes [26], S. Adams [1] and Y. Gao [59]; for results in the setting of CAT(0) spaces, see [114]. The theorem above applies e.g. when G is a semisimple group, or when Γ is a Burger–Mozes group [27], [28], or when Γ is a Kac–Moody group [132], [133].

Problem J. Does there exist a geometric characterisation of the non-vanishing of $\mathrm{H}_{\mathrm{b}}^2(\Gamma, \ell^2(\Gamma))$? Is it a quasi-isometry invariant amongst finitely generated groups?

If this property could be reformulated e.g. in terms of quasi-actions on suitable negatively curved spaces, then such a reformulation could be construed as an analogue of J. Stalling's famous splitting theorem [140]. Indeed, in all known cases our classes in $\mathrm{H}_{\mathrm{b}}^2(\Gamma, \ell^2(\Gamma))$ are in the kernel of the comparison map. Therefore, they appear as *quasifications* of the space $\mathrm{H}^1(\Gamma, \mathbf{R}\Gamma)$ relevant to Stalling's theorem.

Remark 4.3. In a different direction, there is indeed a characterisation of Gromov-hyperbolic groups due to I. Mineyev [109]; the main ingredient therein is the surjectivity of the comparison map in degree two, see M. Gromov [69, 8.3.T] and I. Mineyev [108]. It is unclear whether one could obtain interesting definitions of a *rank* by postulating surjectivity in a given higher degree. Using another type of exotic cohomology, namely ℓ^∞ cohomology, S. Gersten also provided a characterisation of hyperbolicity (see [61] and compare [2]).

4.2. Orbit equivalence. Consider ergodic free measure-preserving actions of a countable group Γ on a probability space (X, μ). The quotient space X/Γ is completely singular, but can nevertheless be investigated by shifting the focus to the type II_1 *measured equivalence relation* $\mathcal{R} \subseteq X \times X$ induced by the action. Accordingly, one calls two actions (by possibly different groups) *orbit equivalent* (OE) if the resulting relations are isomorphic. The Ornstein–Weiss theorem [123] implies that *all*

such action of all amenable countable groups are OE. It is therefore of interest to find obstructions to OE or better yet rigidity results.

The context of Section 4.1 comes into play as follows. Let $\Gamma = \Gamma_1 \times \Gamma_2$ be a product of torsion-free groups with $\mathrm{H}^2_{\mathrm{b}}(\Gamma_i, \ell^2(\Gamma_i)) \neq 0$, e.g. non-elementary hyperbolic groups. Consider a Γ-space (X, μ) that is *irreducible* in the sense that each Γ_i-action is ergodic.

Theorem ([118, 1.6]). *Any Γ-space that is OE to X is actually conjugated to X, possibly twisting the action by an automorphism of Γ.*

Theorem ([118, 1.9]). *If any mildly mixing action of any torsion-free group Λ is OE to X, then Λ is isomorphic to Γ and the actions are conjugated.*

Compare with Hjorth–Kechris [78]. One can also use our techniques to show [118, 1.14]: *There exists a continuum of mutually non weakly isomorphic relations of type II_1 with trivial outer automorphism group.*

5. Randomorphisms

The notion of *randomorphism* between two groups is proposed below. It is closely connected to orbit equivalence and related ideas; thus, not much originality is claimed, except perhaps for the language proposed, which I believe has its own appeal.

5.1. Random maps. The space G^H of all maps $f\colon H \to G$ between the countable groups H, G has a natural structure of Polish space, given by the product uniform structure (with G viewed discrete). Therefore, it makes sense to think of "random maps" $f\colon H \to G$ simply as probability measures on this nice Polish space. To avoid redundancy coming from the free G-action(s), we define the Polish space

$$[H, G] \stackrel{\text{def}}{=} \{f\colon H \to G \ :\ f(e) = e\}$$

(a closed subspace of G^H). There is a natural H-action on $[H, G]$ defined by

$$(h.f)(x) \stackrel{\text{def}}{=} f(xh)f(h)^{-1} \quad \text{for } f \in [H, G],\ h, x \in H.$$

The basic observation is that a homomorphism $H \to G$ is nothing but an H-fixed point for this action.

Definition 5.1. A *randomorphism* from H to G is an H-invariant probability measure on $[H, G]$.

Notice that the subset of injective maps is closed in $[H, G]$. This suggest a naïve notion of injectivity for randomorphisms:

Definition 5.2. A randomorphism is a *randembedding* if it is supported on the injective maps. We say that H is a *random subgroup* of G if it admits a randembedding into G.

Compare with the notion of *placement* proposed by M. Gromov [70, 4.5] and with Y. Shalom's related viewpoint on uniform embeddings [137]. One can verify that a point in $[H, G]$ is almost periodic (i.e. its H-orbit relatively compact) if the corresponding map is Lipschitz in the appropriate sense.

Definition 5.3. A randomorphism, randembedding or random subgroup is *geometric* if the corresponding measure is compactly supported in $[H, G]$.

Problem K. Which groups admit the non-Abelian free group $\mathbf{F}_2$ as a geometric random subgroup?

The following has been proved by D. Gaboriau and independently R. Lyons (private communication); Gaboriau's proof uses percolation techniques, relying among other things on [124] and [77].

Theorem 5.4. *Every non-amenable group admits* $\mathbf{F}_2$ *as a random subgroup.*

5.2. Back to orbit equivalence. If the two countable groups G, H have OE actions as in Section 4.2 on (X, μ) and (Y, ν) respectively, then there is a measure space isomorphism $F\colon X \to Y$ such that $F(H.x) = G.F(x)$ almost everywhere. One can also consider the more general situation where $F(H.x) \subseteq G.F(x)$; for instance, H could be a subgroup of a group having an action orbit equivalent to the G-action. By freeness, there is a measurable map

$$\alpha\colon H \times X \longrightarrow G$$

defined almost everywhere by $F(h.x) = \alpha(h, x).F(x)$. This map is a *cocycle* in that is satisfies

$$\alpha(hk, x) = \alpha(h, k.x)\alpha(k, x) \quad \text{for } h, k \in H,\ \mu - \text{a.e. } x \in X.$$

The cocycle α yields a measurable map $\hat{\alpha}\colon X \to [H, G]$ defined by $\hat{\alpha}(x)(h) = \alpha(h, x)$. This map is H-equivariant, and therefore the measure $\hat{\alpha}_*\mu$ is a randomorphism from H to G.

Observe that regardless of any measure, there is a *tautological cocycle* $E\colon H \times [H, G] \to G$ with respect to the H-action on $[H, G]$ defined by $E(h, f) = f(h)$. In the above construction, the map $\hat{\alpha}$ intertwines the cocycle α to the tautological cocycle. Therefore, $[H, G]$ together with its tautological cocycle has a way to reflecting all OE cocycles within the space of all invariant probability measures on $[H, G]$.

The Ornstein–Weiss theorem [123] shows in particular that every amenable group is a random subgroup of $\mathbf{Z}$. Therefore, one has the following striking "Random Tits Alternative":

Any countable group is either a random subgroup[8] *of* $\mathbf{Z}$ *or has* $\mathbf{F}_2$ *as a random subgroup.*

[8] actually, *measure equivalent* to $\mathbf{Z}$; compare Section 5.5.

The viewpoint of measured relations and OE allows to formulate a related question: Does *any* non-amenable type II_1 relation contains the orbits of a $\mathbf{F}_2$-action? (See Kechris–Miller [96, 28.14] and D. Gaboriau [57, 5.16]). This is only known to hold for relations of non-trivial *cost* (M. Pichot [126], Kechris–Miller [96, 28.8]; see [56] for the notion of cost).

5.3. Modules. Just as a homomorphism $H \to G$ yields a pull-back functor from G-modules to H-modules, we can define pull-backs through randomorphisms:

Definition 5.5. Let V be a coefficient G-module, μ a randomorphism from H to G and $1 \leq p \leq \infty$. The *L^p-pull-back* of V through μ is the Banach space $L^p(\mu, V)$ endowed with the H-action

$$(h.\varphi)(f) \stackrel{\text{def}}{=} f(h^{-1})^{-1}\varphi(h^{-1}.f), \quad h \in H,\ \varphi \in L^p(\mu, V),\ f \in [H, G].$$

We are mostly interested in $p = 2, \infty$. For instance, analysing the case $V = \ell^2(G)$ shows:

Lemma 5.6. *A random subgroup of an amenable group is itself amenable.*

Recall that in the case of injective homomorphisms, the pull-back has an adjoined functor called (*co-*)*induction*. There is again an analogue for randembeddings. Consider G^H with precomposition by right H-translation and postcomposition by right G-translation. This is isomorphic to the E-twisted (cf. [146, p .65]) product H-space $[H, G] \times G$ endowed with an additional G-action by right multiplication. Therefore, it inherits an invariant σ-finite regular Borel measure $\widetilde{\mu}$ defined as the product of μ with the counting measure. The following generalises induction (compare [70, 4.5.C] and [118, 4.1]).

Definition 5.7. Let V be a coefficient H-module, μ a randembedding from H to G and $1 \leq p \leq \infty$. The *L^p-induced module* of V through μ is the Banach space $L^p(\widetilde{\mu}, V)^H$ of H-equivariant maps endowed with the G-action by right translations.

5.4. Application to bounded cohomology. The induction methods used in [118] can be seen to yield the following.

Proposition 5.8. *Let H be a random subgroup of G and V a coefficient H-module. Let W be the L^∞-induced module. For every $n \geq 0$ there is an injection $\mathrm{H}^n_{\mathrm{b}}(H, V) \hookrightarrow \mathrm{H}^n_{\mathrm{b}}(G, W)$.*

(Due to the unwieldy nature of L^∞ spaces, it is essential in [118] to have a similar injectivity statement for the L^p-induced module with $p < \infty$, e.g. $p = 2$. However, the latter is only known to hold when $n \leq 2$.)

Corollary 5.9. *If H is a random subgroup of G, then $\dim^\sharp_{\mathrm{b}}(H) \leq \dim^\sharp_{\mathrm{b}}(G)$.*

Appealing to Theorem 5.4, we conclude:

Corollary 5.10. *No group can have* $\dim_b^\sharp = 1, 2$.

Problem L. Is $\dim_b^\sharp(\mathbf{F}_2)$ infinite? If so, is even $H_b^n(\mathbf{F}_2, \mathbf{R})$ non-zero for all $n \geq 2$?

If $\dim_b^\sharp(\mathbf{F}_2) = \infty$, then it follows that the hierarchy proposed by B. Johnson (cf. beginning of Section 3) collapses completely, since we then have $\dim_b^\sharp(G) = 0$ or ∞ for every group G, according to whether it is amenable or not.

5.5. Categorical approach. As we have seen, orbit equivalences yield randomorphisms. However orbit equivalence is symmetrical; the intuition is that these randomorphism should more precisely be "isomorphisms" in an appropriate category whose morphisms are represented by randomorphisms. A related problem is that the notion of randembedding of Definition 5.2 does not follow the usual categorical pattern that should define mono-randomorphisms. A third issue is that it is unclear when a randomorphism should be considered to be an epimorphism.

In order to address these points, it is necessary to define the composition of two randomorphisms. It seems that there are (at least) two natural composition products:

(i) The independent product. Let G, H, L be countable groups. The composition map

$$[L, H] \times [H, G] \longrightarrow [L, G], \quad (f, f') \longmapsto f' \circ f$$

is continuous. Given probability measures μ, ν on $[L, H]$ respectively $[H, G]$, denote by $\nu \circ \mu$ the image of the product measure $\mu \times \nu$ under this map. Thus $\nu \circ \mu$ is the product of *independently* chosen random maps. If both μ, ν are randomorphisms, then so is $\nu \circ \mu$. The main defect of the independent product is the scarcity of randomorphisms that are invertible for this product.

(ii) The fibred product. Another product has the flavour of groupoids and is defined as follows. Two randomorphisms μ from L to H and ν from H to G are *composable* if there is an isomorphism of Lebesgue spaces

$$F : \big([L, H], \mu\big) \xrightarrow{\cong} \big([H, G], \nu\big)$$

which is equivariant with respect to the tautological cocycle $L \times [L, H] \to H$; that is, $F(\ell.f) = f(\ell).F(f)$ μ-a.e. In that case we define the *fibred product* $\nu \underset{F}{\cdot} \mu$ as the image of μ under the map

$$[L, H] \longrightarrow [L, G], \quad f \longmapsto F(f) \circ f.$$

Both products are a particular case of the construction that associates a randomorphism from L to G to the data of a randomorphism from L to H together with a measurable map from $[L, H]$ to probability measures on $[H, G]$ that is equivariant with respect to the tautological cocycle. In all cases the verification follows from the formula $\ell.(f' \circ f) = (f(\ell).f') \circ (\ell.f)$.

Problem M. (i) Subsume both product in one categorical construction, for instance by defining a suitable equivalence relation on randomorphisms. (ii) Produce an interesting definition of a group of auto-randomorphisms of a given group G. (iii) Reformulate A. Furman's results [51] as a determination of this group for higher rank lattices.

5.6. Random forests. Let G be a countable group and $A \subseteq G$ a non-empty finite subset. Define the compact G-space of unoriented[9], labelled 4-regular graphs as

$$\mathcal{G}_A \stackrel{\text{def}}{=} \left\{(x_{i,g}) \in \prod_{g\in G}(gA)^{\mathbf{Z}/4\mathbf{Z}} : x_{h,i+2} = g \text{ when } h = x_{g,i}\right\}.$$

The topology is the product topology and $k \in G$ acts by $(kx)_{g,i} = k(x_{k^{-1}g,i})$.

Definition 5.11. The space $\mathcal{F}_A$ of 4-forests is the closed G-invariant subspace $\mathcal{F}_A \subseteq \mathcal{G}_A$ of those elements $(x_{i,g})$ for which every connected component is acyclic.

An immediate application of Tarski's theorem on paradoxical decompositions (or of appropriate proofs of it) yields:

Proposition 5.12. *The group G is non-amenable if and only if $\mathcal{F}_A \neq \varnothing$ for A large enough.*

Problem K′. For which groups does there exists an invariant probability measure on $\mathcal{F}_A$ for some A?

Notice that there is a canonical continuous map $\mathcal{F}_A \longrightarrow [\mathbf{F}_2, G]$ which is $\mathbf{F}_2$-equivariant when the left hand side is endowed with the natural $\mathbf{F}_2$-action defined by the labelling of the forests. Moreover, this map ranges in the space of injections. Therefore, groups with a measure as in Problem K′ will have a geometric random free subgroup as in Problem K.

6. Additional questions

Here is the motivation that prompted me to consider randomorphisms to begin with. It is an old observation of J. Dixmier [41] and M. Day [39] that every uniformly bounded representation of an amenable group is *unitarizable*, i.e. conjugated to a unitary representation. The problem of the converse to this statement proposed in J. Dixmier's 1950 article [41] is still open, despite remarkable work most notably by G. Pisier (see [128], [129] and [127]).

It is nevertheless possible to show very explicitly that any group containing $\mathbf{F}_2$ has uniformly bounded representations that are not unitarizable (see e.g. Theorem 2.1 and Lemma 2.7 in [128]). On the other hand, one can also induce such representations from *random subgroups* exactly as explained above and still obtain uniformly bounded representations.

[9] We make the usual convention that an unoriented edge consists of two opposed oriented edges.

Problem N. (i) Is the uniformly bounded representation induced as above non-unitarizable? (ii) Is this true at least for randembeddings of $\mathbf{F}_2$ arising from an invariant measure on the space of forests?

We have seen that the vanishing of H_b^* with general dual coefficients characterises amenability; however, it is not enough to consider trivial coefficients $\mathbf{R}$ even in all degrees. Indeed, refining the method of J. Mather [105], Matsumoto–Morita [107] have proved that the (non-amenable) group of compactly supported homeomorphisms of $\mathbf{R}^n$ has vanishing H_b^* with trivial coefficients. Is the situation nicer for linear groups?

Problem O. (i) Let $\Gamma < \mathrm{GL}_d(\mathbf{C})$ be a subgroup that is not virtually soluble (equivalently, Γ is non-amenable). Is $H_b^n(\Gamma, \mathbf{R})$ non-zero for some $1 < n < d^2$? (ii) If not, is there at least a *separable* dual Banach Γ-module V with $H_b^n(\Gamma, V)$ non-zero for some $n \geq 2$?

One way to approach the problem would be to prove first the conjecture proposed as Problem A′. This done, one needs to deduce geometric finiteness properties from the annihilation of the pull-back $H_b^*(G, \mathbf{R}) \to H_b^*(\Gamma, \mathbf{R})$. The latter step is not at all hopeless, especially in view of the results of B. Klingler for usual cohomology in [98].

One hint to the difficulties that could arise (in addition to Problem A′) is the fact that B. Klingler needs non-trivial coefficient modules. However, as mentioned in the introduction, the analogous statement to Problems A and A′ fails already for unitary representations. Therefore, either one needs to construct cohomology classes for such representations that are genuinely new (as in [25]) and prove results similar to those of [98] for such classes, or one needs to argue that even trivial coefficients suffice because pull-back in bounded cohomology tends to be injective in cases where the usual pull-back is not. (For example, the volume form of the hyperbolic plane restricts non-trivially in H_b^2 to free lattice subgroups of $\mathrm{PSL}_2(\mathbf{R})$, whilst it vanishes when sent to usual cohomology.)

An alternative approach for linear groups could be to consider more generally polynomially bounded cocycles.

Problem P. Let $G = \boldsymbol{G}(k)$ be a simple group of k-rank $r > 0$ over a local field k. Quasify B. Klingler's cocycles [98] in order to obtain new classes in degree $r + 1$ for cohomology with polynomial growth degree $r - 1$ (in a suitable module).

I suspect that $(r + 1)$-cohomology with polynomial growth degree $r - 1$ is indeed the right place to look for "rank r phenomena".

Ch. Bavard proves [6] that a group has non-trivial quasimorphisms if and only if its *stable commutator length* is non-zero.

Problem Q (M. Abért). Let k be a countable field of infinite transcendence degree over its prime field. Does the group $\mathrm{SL}_3(k[X])$ have non-trivial quasi-morphisms?

The interest of $\mathrm{SL}_3(k[X])$ in connection with Ch. Bavard's result is that for fields k as above this group is known to have infinite commutator width, which is a priori not enough to control the stable length.

Finally a question from [113]. For a prime p, denote by $v_p\colon \mathbf{Q}\setminus\{0\}\to\mathbf{Z}$ the p-adic valuation (normalised by $v_p(p^n)=-n$). If q is another prime, define $D_{p,q}\colon \mathbf{Q}\setminus\{0,1\}\to\mathbf{Z}$ by

$$D_{p,q}(x)=v_p(x)v_q(1-x)-v_q(x)v_p(1-x).$$

This function is obviously unbounded; on the other hand, one can form arbitrary linear combinations of such $D_{p,q}$ by varying the primes p,q.

Problem R. Is the function $\sum_{p<q}\alpha_{p,q}D_{p,q}$ unbounded on $\mathbf{Q}\setminus\{0,1\}$ for every family of real numbers $\{\alpha_{p,q}\}$ (unless they are all zero)?

It was observed in [113] that a positive answer would imply $\mathrm{H}^3_{\mathrm{b}}(\mathrm{GL}_2(\mathbf{Q}),\mathbf{R})=0$. Moreover, it follows from the stabilisation results of [113] that the latter vanishing would imply $\mathrm{H}^3_{\mathrm{b}}(\mathrm{GL}_n(\mathbf{Q}_p),\mathbf{R})=0$ for all $n\in\mathbf{N}$ and all primes p.

References

[1] Adams, S., Reduction of cocycles with hyperbolic targets. *Ergodic Theory Dynam. Systems* **16** (6) (1996), 1111–1145.

[2] Allcock, D. J., and Gersten, S. M., A homological characterization of hyperbolic groups. *Invent. Math.* **135** (3) (1999), 723–742.

[3] Anantharaman-Delaroche, C., and Renault, J., *Amenable groupoids*. Monogr. Enseign. Math. 36, L'Enseignement Mathématique, Genève, 2000.

[4] Atiyah, M. F., Elliptic operators, discrete groups and von Neumann algebras. In *Colloque "Analyse et Topologie" en l'Honneur de Henri Cartan*; *Astérisque* **32–33** (1976), 43–72.

[5] Barge, J., and Ghys, É., Surfaces et cohomologie bornée. *Invent. Math.* **92** (3) (1988), 509–526.

[6] Bavard, C., Longueur stable des commutateurs. *Enseign. Math.* (2) **37** (1–2) (1991), 109–150.

[7] Benedetti, R., and Petronio, C., *Lectures on hyperbolic geometry*. Universitext, Springer-Verlag, Berlin 1992.

[8] Bestvina, M., and K. Fujiwara, K., Bounded cohomology of subgroups of mapping class groups. *Geom. Topol.* **6** (2002), 69–89.

[9] Biran, P., Entov, M., and Polterovich, L., Calabi quasimorphisms for the symplectic ball. *Commun. Contemp. Math.* **6** (5) (2004), 793–802.

[10] Bourdon, M., and Pajot, H., Rigidity of quasi-isometries for some hyperbolic buildings. *Comment. Math. Helv.* **75** ((4) (2000), 701–736.

[11] Brooks, R., Some remarks on bounded cohomology. In *Riemann surfaces and related topics: Proceedings of the 1978 Stony Brook Conference*, Ann. of Math. Stud. 97, Princeton University Press, Princeton, N.J., 1981, 53–63.

[12] Brooks, R., and Series, C., Bounded cohomology for surface groups. *Topology* **23** (1) (1984), 29–36.

[13] Bucher-Karlsson, M., Simplicial volume of locally symmetric spaces covered by $SL_3(\mathbb{R})/SO(3)$. Preprint.

[14] Bucher-Karlsson, M., *Characteristic classes and bounded cohomology*. PhD thesis, ETH Zürich, Diss. Nr. 15636, 2004.

[15] Burger, M., and Iozzi, A., Bounded cohomology and deformation rigidity in hyperbolic geometry. Preprint.

[16] Burger, M., and Iozzi, A., Bounded differential forms, generalized Milnor-Wood inequality and an application to deformation rigidity. Preprint, 2005.

[17] Burger, M., and Iozzi, A., A useful formula in bounded cohomology. Preprint, 2005.

[18] Burger, M., and Iozzi, A., Bounded Kähler class rigidity of actions on Hermitian symmetric spaces. *Ann. Sci. École Norm. Sup.* (4) **37** (1) (2004), 77–103.

[19] Burger, M., Iozzi, A., Labourie, F., and Wienhard, A., Maximal representations of surface groups: Symplectic Anosov structures. *Pure Appl. Math. Q.* **1** (3) (2005), 555–601.

[20] Burger, M., Iozzi, A., and Wienhard, A., Hermitian symmetric spaces and Kähler rigidity. *Transform. Groups*, to appear.

[21] Burger, M., Iozzi, A., and Wienhard, A., Tight embeddings. Preprint.

[22] Burger, M., Iozzi, A., and Wienhard, A., Surface group representations with maximal Toledo invariant. *C. R. Math. Acad. Sci. Paris* **336** (5) (2003), 387–390.

[23] Burger, M., and Monod, N., Bounded cohomology of lattices in higher rank Lie groups. *J. Eur. Math. Soc. (JEMS)* **1** (2) (1999), 199–235.

[24] Burger, M., and Monod, N., Continuous bounded cohomology and applications to rigidity theory *Geom. Funct. Anal.* **12** (2) (2002), 219–280.

[25] Burger, M., and Monod, N., On and around the bounded cohomology of SL_2. In *Rigidity in dynamics and geometry*, Springer-Verlag, Berlin 2002, 19–37.

[26] Burger, M., and Mozes, S., CAT(-1)-spaces, divergence groups and their commensurators. *J. Amer. Math. Soc.* **9** (1) (1996), 57–93.

[27] Burger, M., and Mozes, S., Groups acting on trees: from local to global structure. *Inst. Hautes Études Sci. Publ. Math.* **92** (2001), 113–150 , 2000.

[28] Burger, M., and Mozes, S., Lattices in product of trees. *Inst. Hautes Études Sci. Publ. Math.* **(92)** (2000), 151–194.

[29] Burghelea, D., The cyclic homology of the group rings. *Comment. Math. Helv.* **60** (3) (1985), 354–365.

[30] Cannon, J. W., Floyd, W. J., and Parry, W. R., Introductory notes on Richard Thompson's groups. *Enseign. Math.* (2) **42** (3–4) (1996), 215–256.

[31] Cheeger, J., and Gromov, M., L_2-cohomology and group cohomology. *Topology* **25** (2) (1986), 189–215.

[32] Cheeger, J., and Simons, J., Differential characters and geometric invariants. In *Geometry and topology*, Lecture Notes in Math. 1167, Springer-Verlag, Berlin 1985, 50–80.

[33] Christensen, E., Effros, E. G., and Sinclair, A., Completely bounded multilinear maps and C^*-algebraic cohomology. *Invent. Math.* **90** (2) (1987), 279–296.

[34] Clerc, J.-L., and Ørsted, B., The Gromov norm of the Kaehler class and the Maslov index. *Asian J. Math.* **7** (2) (2003), 269–295.

[35] Connell, Christopher, Farb, Benson, The degree theorem in higher rank. *J. Differential Geom.* **65** (1) (2003), 19–59.

[36] Connes, A., On the classification of von Neumann algebras and their automorphisms. In *Convegno sulle Algebre C^* e loro Applicazioni in Fisica Teorica* (INDAM, Rome 1975), Symposia Math. XX, Academic Press, London 1976, 435–478.

[37] Connes, A., Noncommutative differential geometry. *Inst. Hautes Études Sci. Publ. Math.* **62** (1985), 257–360.

[38] Connes, A., and Shlyakhtenko, D., L^2-homology for von Neumann algebras. *J. Reine Angew. Math.* **586** (2005), 125–168.

[39] Day, M. M., Means for the bounded functions and ergodicity of the bounded representations of semi-groups. *Trans. Amer. Math. Soc.* **69** (1950), 276–291.

[40] de la Harpe, P., and Karoubi, M., Représentations approchées d'un groupe dans une algèbre de Banach. *Manuscripta Math.* **22** (3) (1977), 293–310.

[41] Dixmier, J., Les moyennes invariantes dans les semi-groupes et leurs applications. *Acta Sci. Math. Szeged* **12** (1950), 213–227.

[42] Domic, A., and Toledo, D., The Gromov norm of the Kaehler class of symmetric domains. *Math. Ann.* **276** (3) (1987), 425–432.

[43] Dupont, J. L., Bounds for characteristic numbers of flat bundles. In *Algebraic topology, Aarhus 1978*, Lecture Notes in Math. 763, Springer-Verlag, Berlin 1979, 109–119.

[44] Dupont, J. L., and Kamber, F. W., On a generalization of Cheeger-Chern-Simons classes. *Illinois J. Math.* **34** (2) (1990), 221–255.

[45] Entov, M., and Polterovich, L., Calabi quasimorphism and quantum homology. *Internat. Math. Res. Notices* **(30)** (2003), 1635–1676.

[46] Epstein, D. B. A., and Fujiwara, K., The second bounded cohomology of word-hyperbolic groups. *Topology* **36** (6) (1997), 1275–1289.

[47] Eskin, A., and Farb, B., Quasi-flats and rigidity in higher rank symmetric spaces. *J. Amer. Math. Soc.* **10** (3) (1997), 653–692.

[48] Farb, B., and Masur, H., Superrigidity and mapping class groups. *Topology* **37** (6) (1998), 1169–1176.

[49] Fujiwara, K., The second bounded cohomology of a group acting on a Gromov-hyperbolic space. *Proc. London Math. Soc.* (3) **76** (1) (1998), 70–94.

[50] Fujiwara, K., The second bounded cohomology of an amalgamated free product of groups. *Trans. Amer. Math. Soc.* **352** (3) (2000), 1113–1129.

[51] Furman, A., Gromov's measure equivalence and rigidity of higher rank lattices. *Ann. of Math.* (2) **150** (3) (1999), 1059–1081.

[52] Furman, A., Random walks on groups and random transformations. In *Handbook of dynamical systems*, Vol. 1A, North-Holland, Amsterdam 2002, 931–1014.

[53] Furstenberg, H., A Poisson formula for semi-simple Lie groups. *Ann. of Math.* (2) **77** (1963), 335–386.

[54] Furstenberg, H., Random walks and discrete subgroups of Lie groups. In *Advances in Probability and Related Topics*, Vol. 1, Dekker, New York 1971, 1–63..

[55] Furstenberg, H., Rigidity and cocycles for ergodic actions of semisimple Lie groups (after G. A. Margulis and R. Zimmer). In *Bourbaki Seminar, Vol. 1979/80*, Lecture Notes in Math. 842, Springer-Verlag, Berlin 1981, 273–292.

[56] Gaboriau, D., Coût des relations d'équivalence et des groupes. *Invent. Math.* **139** (1) (2000), 41–98.

[57] Gaboriau, D., Arbres, groupes, quotients. Mémoire d'habilitation, ÉNS-Lyon, 2002.

[58] Gambaudo, J.-M., and Ghys, É., Commutators and diffeomorphisms of surfaces. *Ergodic Theory Dynam. Systems* **24** (5) (2004), 1591–1617.

[59] Gao, Y., Superrigidity for homomorphisms into isometry groups of CAT(-1) spaces. *Transform. Groups* **2** (3) (1997), 289–323.

[60] Garnett, L., Foliations, the ergodic theorem and Brownian motion. *J. Funct. Anal.* **51** (3) (1983), 285–311.

[61] Gersten, S. M., A cohomological characterization of hyperbolic groups. Unpublished, http://www.math.utah.edu/∼gersten.

[62] Gevirtz, J., Stability of isometries on Banach spaces. *Proc. Amer. Math. Soc.* **89** (4) (1983), 633–636.

[63] Ghys, É., Groupes d'homéomorphismes du cercle et cohomologie bornée. In *The Lefschetz centennial conference*, Part III, Amer. Math. Soc., Providence, RI, 1987, 81–106.

[64] Ghys, É., and de la Harpe, P. (eds.), *Sur les groupes hyperboliques d'après Mikhael Gromov*. Progr. Math. 83, Birkhäuser, Boston 1990.

[65] Goldman, W. M., and Millson, J. J., Local rigidity of discrete groups acting on complex hyperbolic space. *Invent. Math.* **88** (1987), 495–520.

[66] Grigorchuk, R. I., Some results on bounded cohomology. In *Combinatorial and geometric group theory*, London Math. Soc. Lecture Note Ser. 204, Cambridge University Press, Cambridge, 1995, 111–163.

[67] Gromov, M., Volume and bounded cohomology. *Inst. Hautes Études Sci. Publ. Math.* **(56)** (1982), 5–99.

[68] Gromov, M., Infinite groups as geometric objects. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 1, PWN, Warsaw 1984, 385–392.

[69] Gromov, M., Hyperbolic groups. In *Essays in group theory*, Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–263.

[70] Gromov, M., Rigid transformations groups. In *Géométrie différentielle* (Paris, 1986), Travaux en Cours 33, Hermann, Paris 1988, 65–139.

[71] Gromov, M., Asymptotic invariants of infinite groups. In *Geometric group theory* (Sussex, 1991), Vol. 2, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge, 1993, 1–295.

[72] Grothendieck, A., Produits tensoriels topologiques et espaces nucléaires. *Mem. Amer. Math. Soc.* **1955** (16) (1955), 140pp.

[73] Gruber, P. M., Stability of isometries. *Trans. Amer. Math. Soc.* **245** (1978), 263–277.

[74] Hamenstädt, U., Bounded cohomology and isometry groups of hyperbolic spaces. Preprint.

[75] Hamenstädt, U., Isometry groups of proper hyperbolic spaces. Preprint.

[76] Helemskiĭ, A. Y., *The homology of Banach and topological algebras*. Math. Appl. (Soviet Ser.) 41, Kluwer Academic Publishers Group, Dordrecht 1989.

[77] Hjorth, G., A lemma for cost attained. Preprint.

[78] Hjorth, G., and Kechris, A. S., Rigidity theorems for actions of product groups and countable Borel equivalence relations. *Mem. Amer. Math. Soc.* **177** (833) (2005), 109pp.

[79] Hopf, E., Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. *Ber. Verh. Sächs. Akad. Wiss. Leipzig* **91** (1939), 261–304.

[80] Hyers, D. H., On the stability of the linear functional equation. *Proc. Nat. Acad. Sci. U. S. A.* **27** (1941), 222–224.

[81] Hyers, D. H., and Ulam, S. M., On approximate isometries. *Bull. Amer. Math. Soc.* **51** (1945), 288–292.

[82] Hyers, D. H., and Ulam, S. M., Approximately convex functions. *Proc. Amer. Math. Soc.* **3** (1952), 821–828.

[83] Hyers, D. H., and Ulam, S. M., On the stability of differential expressions. *Math. Mag.* **28** (1954), 59–64.

[84] Iozzi, A., Bounded cohomology, boundary maps, and rigidity of representations into $\mathrm{Homeo}^+(\mathbf{S^1})$ and $\mathrm{SU}(1,n)$. In *Rigidity in Dynamics and Geometry*. Springer-Verlag, Berlin 2002.

[85] Ivanov, N. V., Foundations of the theory of bounded cohomology. *J. Soviet Math.* **37** (1987), 1090–1115.

[86] Ivanov, N. V., and Turaev, V. G., The canonical cocycle for the Euler class of a flat vector bundle. *Dokl. Akad. Nauk SSSR* **265** (3) (1982), 521–524.

[87] Ji, R., Nilpotency of Connes' periodicity operator and the idempotent conjectures. *K-Theory* **9** (1) (1995), 59–76.

[88] Johnson, B. E., Approximate diagonals and cohomology of certain annihilator Banach algebras. *Amer. J. Math.* **94** (1972), 685–698.

[89] Johnson, B. E., Cohomology in Banach algebras. *Mem. Amer. Math. Soc.* **127** (1972), 96pp.

[90] Kaĭmanovich, V. A., The Poisson formula for groups with hyperbolic properties. *Ann. of Math.* (2) **152** (3) (2000), 659–692.

[91] Kaĭmanovich, V. A., Double ergodicity of the Poisson boundary and applications to bounded cohomology. *Geom. Funct. Anal.* **13** (4) (2003), 852–861.

[92] Kaĭmanovich, Vadim A. Amenability and the Liouville property. Probability in mathematics. *Israel J. Math.* **149** (2005), 45–85.

[93] Kaĭmanovich, V. A., and Masur, H., The Poisson boundary of the mapping class group. *Invent. Math.* **125** (2) (1996), 221–264.

[94] Kaĭmanovich, V. A., and Vershik, A. M., Random walks on discrete groups: boundary and entropy. *Ann. Probab.* **11** (3) (1983), 457–490.

[95] Kazhdan, D., On ε-representations. *Israel J. Math.* **43** (4) (1982), 315–323.

[96] Kechris, A. S., and Miller, B. D., *Topics in orbit equivalence*. Lecture Notes in Math. 1852, Springer-Verlag, Berlin 2004.

[97] Kleiner, B., and Leeb, B., Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings. *Inst. Hautes Études Sci. Publ. Math.* **(86)** (1997), 115–197.

[98] Klingler, B., Volumes des représentations sur un corps local. *Geom. Funct. Anal.* **13** (5) (2003), 1120–1160.

[99] Kuessner, T., *Relative simplicial volume*. PhD thesis, Universität Tübingen, 2001.

[100] Lafont, J.-F., and Schmidt, B., Simplicial volume of closed locally symmetric spaces of non-compact type. *Acta Math.*, to appear.

[101] Lück, W., Dimension theory of arbitrary modules over finite von Neumann algebras and L^2-Betti numbers. I. Foundations. *J. Reine Angew. Math.* **495** (1998), 135–162.

[102] Manning, J. F., Geometry of pseudocharacters. *Geom. Topol.* **9** (2005), 1147–1185..

[103] Margulis, G. A., On the decomposition of discrete subgroups into amalgams. *Selecta Math. Soviet.* **1** (2) (1981), 197–213.

[104] Margulis, G. A., *Discrete subgroups of semisimple Lie groups*. Ergeb. Math. Grenzgeb. 17, Springer-Verlag, Berlin 1991.

[105] Mather, J. N., The vanishing of the homology of certain groups of homeomorphisms. *Topology* **10** (1971), 297–298.

[106] Matsumoto, S., Some remarks on foliated S^1 bundles. *Invent. Math.* **90** (1987), 343–358.

[107] Matsumoto, S., and Morita, S., Bounded cohomology of certain groups of homeomorphisms. *Proc. Amer. Math. Soc.* **94** (3) (1985), 539–544.

[108] Mineyev, I., Straightening and bounded cohomology of hyperbolic groups. *Geom. Funct. Anal.* **11** (4) (2001), 807–839.

[109] Mineyev, I., Bounded cohomology characterizes hyperbolic groups. *Q. J. Math.* **53** (1) (2002), 59–73.

[110] Mineyev, I., Monod, N., and Shalom, Y., Ideal bicombings for hyperbolic groups and applications. *Topology* **43** (6) (2004), 1319–1344.

[111] Mitsumatsu, Y., Bounded cohomology and l^1-homology of surfaces. *Topology* **23** (4) (1984), 465–471.

[112] Monod, N., *Continuous bounded cohomology of locally compact groups*. Lecture Notes in Math. 1758, Springer-Verlag, Berlin 2001.

[113] Monod, N., Stabilization for SL_n in bounded cohomology. In *Discrete geometric analysis*, Contemp. Math. 347, Amer. Math. Soc., Providence, RI, 2004, 191–202.

[114] Monod, N., Superrigidity for irreducible lattices and geometric splitting. *J. Amer. Math. Soc.*, to appear 2006.

[115] Monod, N., and Rémy, B., Boundedly generated groups with pseudocharacter(s). Appendix to: J. F. Manning, Quasi-actions on trees and Property (QFA), *J. London Math. Soc.* **73** (1) (2006), 104–108.

[116] Monod, N., and Shalom, Y., Negative curvature from a cohomological viewpoint and cocycle superrigidity. *C. R. Acad. Sci. Paris Sér. I Math.* **337** (10) (2003), 635–638.

[117] Monod, Nicolas, and Shalom, Yehuda, Cocycle superrigidity and bounded cohomology for negatively curved spaces. *J. Differential Geom.* **67** (3) (2004), 395–455.

[118] Monod, N., and Shalom, Y., Orbit equivalence rigidity and bounded cohomology. *Ann. of Math.*, to appear, 2006.

[119] Mostow, G. D., The rigidity of locally symmetric spaces. In *Actes du Congrès International des Mathématiciens* (Nice, 1970), Tome 2, Gauthier-Villars, Paris 1971, 187–197.

[120] Noskov, G. A., Bounded cohomology of discrete groups with coefficients. *Leningr. Math. J.* **2** (5) (1991), 1067–1084.

[121] Noskov, G. A., The Hochschild-Serre spectral sequence for bounded cohomology. In *Proceedings of the International Conference on Algebra*, Part 1, Contemp. Math. 131, Amer. Math. Soc., Providence, RI, 1992, 613–629.

[122] Omladič, M., and Šemrl, P., On nonlinear perturbations of isometries. *Math. Ann.* **303** (4) (1995), 617–628.

[123] Ornstein, D. S., and Weiss, B., Ergodic theory of amenable group actions. I. The Rohlin lemma. *Bull. Amer. Math. Soc.* **2** (1) (1980), 161–164.

[124] Pak, I., and Smirnova-Nagnibeda, T., On non-uniqueness of percolation on nonamenable Cayley graphs. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (6) (2000), 495–500.

[125] Pansu, P., Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rang un. *Ann. of Math.* (2) **129** (1) (1989), 1–60.

[126] Pichot, M., Quasi-périodicité et théorie de la mesure. PhD thesis, ÉNS-Lyon, 2005.

[127] Pisier, G., Simultaneous similarity, bounded generation and amenability. Preprint.

[128] Pisier, G., *Similarity problems and completely bounded maps*, Lecture Notes in Math. 1618, expanded edition, Springer-Verlag, Berlin 2001.

[129] Pisier, G., Are unitarizable groups amenable? In *Infinite groups: geometric, combinatorial and dynamical aspects*, Progr. Math. 248, Birkhäuser, Basel 2005, 323–362.

[130] Popa, S., Correspondences. INCREST preprint, 1986.

[131] Py, P., Quasi-morphisme de Calabi sur les surfaces de genre supérieur. *C. R. Math. Acad. Sci. Paris* **341** (1) (2005), 29–34.

[132] Rémy, B., Groupes de Kac-Moody déployés et presque déployés. *Astérisque* **277** (2002), 348pp.

[133] Rémy, B., Kac-Moody groups as discrete groups. To appear in the proceedings of *Geometric Group Theory* (2002), Hindustan Book Agency.

[134] Savage, R. P., Jr., The space of positive definite matrices and Gromov's invariant. *Trans. Amer. Math. Soc.* **274** (1) (1982), 239–263.

[135] Sela, Z., Uniform embeddings of hyperbolic groups in Hilbert spaces. *Israel J. Math.* **80** (1–2) (1992), 171–181.

[136] Shalom, Y., Rigidity of commensurators and irreducible lattices. *Invent. Math.* **141** (1) (2000), 1–54.

[137] Shalom, Y., Harmonic analysis, cohomology, and the large-scale geometry of amenable groups. *Acta Math.* **192** (2) (2004), 119–185.

[138] Smillie, J., An obstruction to the existence of affine structures. *Invent. Math.* **64** (3) (1981), 411–415.

[139] Soma, T., The zero-norm subspace of bounded cohomology. *Comment. Math. Helv.* **72** (4) (1997), 582–592.

[140] Stallings, J., On torsion-free groups with infinitely many ends. *Ann. of Math.* (2) **88** (1968), 312–334.

[141] Sullivan, D., A generalization of Milnor's inequality concerning affine foliations and affine manifolds. *Comment. Math. Helv.* **51** (2) (1976), 183–189.

[142] Thurston, W. P., *Three-dimensional geometry and topology*. Princeton Math. Ser. 35, Princeton University Press, Princeton, NJ, 1997.

[143] Wienhard, A., Bounded cohomology and geometry. PhD thesis, Universität Bonn, 2004.

[144] Zimmer, R. J., Amenable ergodic actions, hyperfinite factors, and Poincaré flows. *Bull. Amer. Math. Soc.* **83** (5) (1977), 1078–1080.

[145] Zimmer, R. J., Amenable ergodic group actions and an application to Poisson boundaries of random walks. *J. Functional Analysis* **27** (3) (1978), 350–372.

[146] Zimmer, R. J., *Ergodic theory and semisimple groups*. Monogr. Math. 81, Birkhäuser Verlag, Basel 1984.

Université de Genève, 2-4, rue du Lièvre, 1211 Genève, Switzerland
E-mail: nicolas.monod@unige.ch

Fibration de Hitchin et structure endoscopique de la formule des traces

Bao-Châu Ngô

Abstract. The Hitchin fibration is a well suited tool to study the geometric side of the trace formula for Lie algebra from the point of view of moduli spaces of vector bundles over a curve. The endoscopy groups appear naturally when we decompose the cohomology of the Hitchin fibration by its natural symmetries. Following this dictionary, we can formulate a global and geometric version of Langlands–Shelstad's fundamental lemma. This conjecture has been proved in the case of unitary groups in a joint work with G. Laumon.

Résumé. La fibration de Hitchin fournit un outil adapté pour explorer le côté géométrique de la formule des traces pour l'algèbre de Lie du point de vue des espaces de module des fibrés sur une courbe. Les groupes d'endoscopie apparaissent naturellement quand on cherche à décomposer la cohomologie de la fibration de Hitchin par ses symétries naturelles. En poursuivant ce dictionnaire, on peut formuler une version globale et géométrique du lemme fondamental de Langlands et Shelstad. Cette conjecture a été démontrée dans le cas particulier des groupes unitaires dans un travail en commun avec G. Laumon.

Mathematics Subject Classification (2000). Primary 14H60, 11F72 ; Secondary 22E35.

Keywords. Trace formula, endoscopy groups, fundamental lemma, moduli spaces of vector bundles.

Mots-clés. Formule des traces, groupes endoscopiques, lemme fondamental, espace de module des fibrés.

1. Commentaires historiques

La stabilisation de la formule des traces est l'un des objectifs de la théorie des représentations automorphes. On ne saurait mieux l'introduire que Langlands l'a fait dans les premières lignes de [15] :

En principe la formule des traces exprime la trace comme une somme sur les classes de conjugaison d'un groupe $G(F)$, F étant un corps global et G un groupe réductif. Donc si l'on veut par exemple comparer la trace pour un groupe quasi-déployé G^* et une forme intérieure G comme on l'a fait dans [13] *il faut trouver une application naturelle de l'ensemble des classes de conjugaison de $G(F)$ dans celui des classes de $G^*(F)$, ce qui est en général impossible. En revanche, si on passe*

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

aux classes de conjugaison stable une telle application existe et facile à définir. Par conséquent pour imiter les méthodes de [13] *il faut d'abord trouver une formule des traces qui s'exprime comme somme sur les classes de conjugaison stable ...*

Les travaux de Langlands et Kottwitz ont fait surgir les groupes endoscopiques de la structure interne de la somme sur les classes de conjugaison elliptiques de $G(F)$. Pour stabiliser la partie elliptique de la formule des traces, il reste à démontrer la conjecture de transfert et le lemme fondamental. Ces conjectures consistent en gros en une comparaison d'intégrales orbitales sur G et sur un groupe endoscopique H. Les travaux de Langlands et Kottwitz sur la partie elliptique ont été considérablement généralisés par Arthur à toute la formule des traces [1].

Les méthodes dites élémentaires ont permis d'établir ces conjectures dans les cas suivants : SL(2) par Labesse et Langlands [14], U(3) par Rogawski [23], Sp(4) par Hales [9], Weissauer [28] et SL(n) par Waldspurger [24]. De plus, Waldspurger a démontré que dans le cas des algèbres de Lie, le lemme fondamental implique la conjecture de transfert [26].

Plus récemment, les travaux de Goresky, Kottwitz et MacPherson [7], de Laumon [17], [18], de Laumon et moi-même [20] donnent un espoir qu'on pourrait démontrer le lemme fondamental en général par des méthodes géométriques. Nous renvoyons à l'exposé Bourbaki de J.-F. Dat [4] et au rapport de Laumon dans ces volumes pour un survol de l'approche géométrique du lemme fondamental.

L'objet du présent rapport est de décrire géométriquement la structure endoscopique de la partie anisotrope la formule des traces [15], [12] suivant [21]. Cette étape se trouve historiquement en amont du lemme fondamental. Du point de vue géométrique, elle permet d'interpoler le lemme fondamental dans une déformation naturelle et d'utiliser des arguments en famille comme la notion de faisceaux pervers purs qui sont déterminants dans [20].

Notre interprétation géométrique est fondée sur la fibration de Hitchin. C'est une variante globale des fibres de Springer affines que Goresky, Kottwitz et MacPherson ont utilisées pour interpréter les intégrales orbitales locales. Les fibres de la fibration de Hitchin pour GL(n) sont des jacobiennes compactifiées de même type que celles qui sont construites par Laumon dans [17].

L'inconvénient de l'approche géométrique est qu'elle s'adapte bien mal à la caractéristique zéro. Heureusement, grâce aux travaux de Waldspurger [27] et de Cluckers–Loeser [3], le lemme fondamental en caractéristique zéro peut se déduire du lemme fondamental en caractéristique positive.

2. Fibration de Hitchin

Soit X une courbe projective lisse sur un corps k. Soit D un diviseur de X de degré $\deg(D) > 2g - 2$ où $g > 2$ est le genre de X. Soit G un groupe réductif sur k ou plus généralement un schéma en groupes réductifs sur X. Pour un triplet (X, D, G)

donné, on considère l'espace de module $\mathcal{M}$ des couples (E, ϕ) où E est un G-torseur sur X et où ϕ est une section

$$\phi \in \mathrm{H}^0(X,\ \mathrm{ad}(E)(D))$$

où $\mathrm{ad}(E)$ est le fibré vectoriel qui se déduit du G-torseur E et de la représentation adjointe de G et où on a noté $V(D) = V \otimes_{\mathcal{O}_X} \mathcal{O}_X(D)$ pour tout fibré vectoriel V sur X. Il est plus agréable de considérer le champ algébrique $\mathcal{M}$, voir [19], plutôt que l'espace de module grossier associé à l'ouvert semi-stable de ce champ. La structure symplectique sur $\mathcal{M}$, qui existe naturellement dans le cas où D est le diviseur canonique de X, ne semble pas jouer de rôle dans notre problème.

Dans le cas où $G = \mathrm{GL}(n)$, la donnée de E consiste en la donnée d'un fibré vectoriel V de rang n sur X et celle de ϕ est équivalente à la donnée d'un endomorphisme tordu $\phi\colon V \to V(D)$. On renvoie à [10] pour une description similaire dans le cas où G est un groupe classique.

Soit $\mathfrak{g}$ l'algèbre de Lie de G, $k[\mathfrak{g}]$ l'anneau des fonctions polynômiales sur $\mathfrak{g}$. Soit $\mathfrak{t}$ l'algèbre de Lie d'un tore maximal T de G et soit W le groupe de Weyl. Si k est un corps de caractéristique nulle ou grande par rapport à $\mathfrak{g}$, on sait d'après Chevalley et Kostant que l'anneau des invariants $k[\mathfrak{g}]^G = k[\mathfrak{t}]^W$ est un anneau des polynômes $k[u_1, \ldots, u_r]$ où r est le rang de G et où $u_1, \ldots, u_r$ sont des polynômes homogènes de degrés $d_1, \ldots, d_r$. Notons

$$\mathfrak{t}/W = \mathrm{Spec}(k[\mathfrak{t}]^W) = \mathrm{Spec}(k[u_1, \ldots, u_r]).$$

L'inclusion $k[\mathfrak{g}]^G \to k[\mathfrak{g}]$ définit le morphisme caractéristique de Chevalley

$$\chi\colon \mathfrak{g} \to \mathfrak{t}/W \tag{1}$$

qui généralise la construction du polynôme caractéristique dans le cas $\mathrm{GL}(n)$. En appliquant χ à la section $\phi \in \mathrm{H}^0(X, \mathrm{ad}(E)(D))$, on trouve une section

$$a \in \bigoplus_{i=1}^{r} \mathrm{H}^0(X, \mathcal{O}_X(d_i D)).$$

Il revient au même de dire que a est une section globale $(\mathfrak{t}/W) \times^{\mathbb{G}_m} L_D$ où L_D est le $\mathbb{G}_m$-torseur sur X associé au fibré en droites $\mathcal{O}_X(D)$.

On obtient ainsi une fibration

$$m\colon \mathcal{M} \to \mathcal{A}$$

où $\mathcal{A} = \bigoplus_{i=1}^{r} \mathrm{H}^0(X, \mathcal{O}_X(d_i D))$ est un espace affine sur k dont la dimension dépend du triplet (X, D, G). Dans le cas $G = \mathrm{GL}(n)$, pour tout point (V, ϕ) de $\mathcal{M}$, on a $m(V, \phi) = (a_1, \ldots, a_n)$ où $a_i \in \mathrm{H}^0(X, \mathcal{O}_X(iD))$ est la trace de $\wedge^i \phi\colon \wedge^i V \to \wedge^i V(iD)$.

Supposons que $k = \mathbb{F}_q$ est un corps fini. En suivant le comptage de Weil des fibrés vectoriels sur une courbe, on peut exprimer formellement le nombre pondéré des points de $\mathcal{M}(\mathbb{F}_q)$

$$|\mathcal{M}(\mathbb{F}_q)| = \sum_{(E,\phi)\in\mathcal{M}(\mathbb{F}_q)} \frac{1}{\mathrm{Aut}(E,\phi)}$$

en termes d'intégrales adéliques. Soient F le corps des fonctions rationnelles sur X, F_v le complété de F en un point fermé $v \in |X|$, $\mathcal{O}_v$ l'anneau des entiers de F_v et $\mathbb{A}_F$ l'anneau des adèles de F. On a alors

$$|\mathcal{M}(\mathbb{F}_q)| = \sum_{\xi\in \ker^1(F,G)} \sum_{\gamma\in\mathfrak{g}^\xi(F)/\sim} O_\gamma(1_D) \tag{2}$$

où

- $\ker^1(F, G)$ est l'ensemble des classes d'isomorphisme des G-torseurs sur F, triviaux localement sur chaque F_v ;
- $\mathfrak{g}^\xi$ est la forme de $\mathfrak{g}$ sur F définie par ξ ;
- γ parcourt l'ensemble des classes de conjugaison de $\mathfrak{g}^\xi(F)$;
- $O_\gamma(1_D)$ est l'intégrale orbitale globale

$$O_\gamma(1_D) = \int_{G_\gamma(F)\backslash G(\mathbb{A}_F)} 1_D(\mathrm{ad}(g)^{-1}\gamma)\, dg$$

de la fonction

$$1_D = \bigotimes_{v\in|X|} 1_{D_v},$$

1_{D_v} étant la fonction caractéristique du compact ouvert $\varpi^{-d_v}\mathfrak{g}(\mathcal{O}_v)$ de $\mathfrak{g}(F_v)$, les entiers d_v étant définis par la formule $D = \sum_{v\in|X|} d_v v$;

- dg est la mesure de Haar normalisée de $G(\mathbb{A})$ de telle façon que $G(\mathcal{O}_\mathbb{A})$ ait volume 1.

L'égalité ci-dessus n'a pas de sens numérique, ses deux membres n'étant pas finis, mais elle peut s'interpréter en termes d'équivalence de catégories. L'égalité heuristique peut néanmoins servir comme un bon guide. On reconnaît dans le membre de droite l'expression formelle du côté géométrique de la formule des traces pour l'algèbre de Lie. Cette égalité formelle est donc le début d'un dictionnaire que nous allons poursuivre.

Pour $a \in \mathcal{A}(\mathbb{F}_q)$, on note $\mathcal{M}_a = m^{-1}(a)$ la fibre de m en a. Le point a peut être vu comme un élément de $(\mathfrak{t}/W)(F)$ vérifiant une condition d'intégralité par rapport à D. Supposons désormais que a est une caractéristique semi-simple régulière c'est-à-dire il correspond à un F-point du l'ouvert $\mathfrak{t}/W$ où le morphisme $\mathfrak{t} \to \mathfrak{t}/W$ est étale. Cette condition semi-simple régulière définit un ouvert non vide $\mathcal{A}^\heartsuit$ de $\mathcal{A}$. Pour tout

$a \in \mathcal{A}(\mathbb{F}_q)$, on a formellement

$$|\mathcal{M}_a(\mathbb{F}_q)| = \sum_{\xi \in \ker^1(F,G)} \sum_{\substack{\gamma \in \mathfrak{g}^\xi(F)/\sim \\ \chi(\gamma)=a}} O_\gamma(1_D). \tag{3}$$

Lorsque $a \in \mathcal{A}^\heartsuit(\mathbb{F}_q)$ la seconde somme s'étend donc sur une classe de conjugaison stable globale dans $\mathfrak{g}^\xi(F)$. Ainsi la fibration de Hitchin correspond essentiellement au découpage de la formule des traces en des classes de conjugaison stable globale.

Notons aussi que d'après la formule des traces de Grothendieck–Lefschetz, on a aussi formellement

$$|\mathcal{M}_a(\mathbb{F}_q)| = \mathrm{Tr}(\mathrm{F}_q, (m_*\mathbb{Q}_\ell)_a) \tag{4}$$

où $m_*\mathbb{Q}_\ell$ est l'image directe dérivée du faisceau constant $\mathbb{Q}_\ell$ et où F_q est l'endomorphisme de Frobenius géométrique. Ainsi le complexe $m_*\mathbb{Q}_\ell$ interpole les sommes (3) dépendant du paramètre a.

Notons que lorsque a appartient à l'ouvert anisotrope $\mathcal{A}^{\mathrm{ani}}$ de $\mathcal{A}^\heartsuit$, les deux côtés de (3) sont des sommes finies et on a dans ce cas une égalité numérique au lieu d'une équivalence de catégories. Nous envoyons à [21] pour la définition précise des ouverts $\mathcal{A}^{\mathrm{ani}} \subset \mathcal{A}^\heartsuit$ de $\mathcal{A}$. Pour les lecteurs familiers avec les courbes spectrales de [10], mentionnons que dans le cas $G = \mathrm{GL}(n)$ et la caractéristique p est plus grande que n, $\mathcal{A}^\heartsuit$ consiste en les caractéristiques a telles que la courbe spectrale Y_a est réduite. L'ouvert $\mathcal{A}^{\mathrm{ani}}$ est non vide seulement pour les G semi-simples et contient alors les $a \in \mathcal{A}$ telle que la courbe spectrale Y_a est irréductible dans le cas $\mathrm{SL}(n)$. Pour les groupes classiques, $a \in \mathcal{A}^{\mathrm{ani}}$ si l'involution naturelle du groupe classique agit trivialement sur l'ensemble des composantes irréductibles de la courbe spectrale définie dans [9]. Les formules (2) et (3) ont un sens numérique quand nous nous limitons à la partie anisotrope.

Théorème 2.1. *La fibration de Hitchin $m\colon \mathcal{M} \to \mathcal{A}$ est lisse sur l'ouvert $\mathcal{A}^\heartsuit$ et est propre sur l'ouvert $\mathcal{A}^{\mathrm{ani}}$.*

Les arguments nécessaires pour démontrer ce théorème étaient déjà dans la littérature, voir notamment [2] et [6], seules les définitions des ouverts $\mathcal{A}^\heartsuit$ et $\mathcal{A}^{\mathrm{ani}}$ ont apparu plus tard dans [21]. L'énoncé de lissité a été démontré dans [21], celui de propreté dans le cas particulier du groupe unitaire dans [20].

3. Stabilisation de la partie anisotrope

La partie anisotropique de la somme (3) peut être transformée de la même façon que la stabilisation de la partie elliptique de la formule des traces suivant Langlands et Kottwitz [15] et [12]. À la différence de la formule des traces, le comptage de points de $\mathcal{M}$, tout comme celui des points des variétés de Shimura et de Drinfeld, comporte en plus une première sommation sur $\ker^1(F, G)$ qui, en fait, simplifie la stabilisation.

Supposons que a est semi-simple régulier. Donnons-nous un élément $\gamma_0 \in \mathfrak{g}(F)$ d'image $\chi(\gamma_0) = a$. Le centralisateur I_{γ_0} est un tore qui ne dépend pas du choix de γ_0. On le notera I_a et on le supposera *anisotrope*. Le tore dual $\hat{I}_a$ est muni d'une action finie de $\Gamma = \mathrm{Gal}(\overline{F}/F)$ telle que $\hat{I}_a^{\Gamma}$ est fini.

Pour tout $\xi \in \ker^1(F, G)$, les classes de conjugaison $\gamma \in \mathfrak{g}^{\xi}(F)$ telles que $\chi(\gamma) = a$ sont en bijection avec les classes de cohomologie

$$\alpha = \mathrm{inv}(\gamma_0, \gamma) \in \mathrm{H}^1(F, I_a)$$

dont l'image dans $\mathrm{H}^1(F, G)$ est l'élément ξ. Ainsi l'ensemble des paires (ξ, γ) de la somme (3) où $\xi \in \ker^1(F, G)$ et γ est une classe de conjugaison de $\mathfrak{g}^{\xi}(F)$ d'image $a \in \mathscr{A}^{\mathrm{ani}}(\mathbb{F}_q)$ est en bijection avec

$$\ker[\mathrm{H}^1(F, I_a) \to \bigoplus_{v \in |X|} \mathrm{H}^1(F_v, G)].$$

Pour toute place $v \in |X|$, donner une classe de conjugaison $\gamma_v \in \mathfrak{g}(F_v)$ telle que $\chi(\gamma_v) = a$ revient à donner son invariant

$$\alpha_v = \mathrm{inv}_v(\gamma_0, \gamma_v) \in \ker[\mathrm{H}^1(F_v, I_a) \to \mathrm{H}^1(F_v, G)].$$

D'après Kottwitz, ce groupe peut être décrit en termes des groupes duaux de la façon suivante. Soit Γ_v le groupe de Galois local $\mathrm{Gal}(\overline{F}_v/F_v)$. D'après la dualité de Nakayama, donner un élément $\alpha_v \in \mathrm{H}^1(F_v, I_a)$ revient à donner un caractère d'ordre fini $\alpha_v : \hat{I}_a^{\Gamma_v} \to \mathbb{C}^{\times}$. Pour que l'élément α_v ait l'image triviale dans $\mathrm{H}^1(F_v, G)$, il faut et il suffit que la restriction du caractère α_v à $Z_{\hat{G}}^{\Gamma_v}$ est triviale où $Z_{\hat{G}}$ est le centre du groupe dual $\hat{G}$.

Pour qu'une collection de classes de conjugaison $(\gamma_v)_{v \in |X|}$ de $\mathfrak{g}(F_v)$ avec $\chi(\gamma_v) = a$ provienne d'une paire (ξ, γ) de la somme (3), il faut et il suffit que $\gamma_v = \gamma_0$ pour presque tout v et que

$$\sum_{v \in |X|} \alpha_v|_{\hat{I}_a^{\Gamma}} = 0 \tag{5}$$

où $\alpha_v = \mathrm{inv}_v(\gamma_0, \gamma_v)$. Si c'est le cas, le nombre des paires (ξ, γ) qui s'envoie sur cette collection $(\gamma_v)_{v \in |X|}$ est égal au cardinal du groupe

$$\ker^1(F, I_a) = \ker[\mathrm{H}^1(F, I_a) \to \bigoplus_{v \in |X|} \mathrm{H}^1(F_v, I_a)].$$

La somme (3) se récrit comme suit

$$|\ker^1(F, I_a)|\, \tau(I_a) \sum_{(\gamma_v)_{v \in |X|}} \prod_v O_{\gamma_v}(1_{D_v}) \tag{6}$$

où les γ_v sont des classes de conjugaison de $\mathfrak{g}(F_v)$ vérifiant l'équation (5). En mettant en facteur le nombre de Tamagawa $\tau(I_a)$, on trouve une somme de produits d'intégrales orbitales locales $\prod_v O_{\gamma_v}(1_{D_v})$ au lieu des intégrale globales $O_\gamma(1_D)$.

En appliquant la formule d'Ono [22]

$$|\ker^1(F, I_a)|\,\tau(I_a) = |\pi_0(\hat{I}_a^\Gamma)|,$$

la somme (3) devient

$$|\pi_0(\hat{I}_a^\Gamma)| \sum_{(\gamma_v)_{v\in|X|}} \prod_v O_{\gamma_v}(1_{D_v}) \tag{7}$$

où les (γ_v) vérifie la condition (5). Notons qu'avec l'hypothèse I_a anisotrope, le groupe $\hat{I}_a$ est un groupe fini de sorte que $\pi_0(\hat{I}_a^\Gamma) = \hat{I}_a^\Gamma$.

En utilisant la transformation de Fourier sur le groupe fini $\hat{I}_a^\Gamma$, la somme (3) devient

$$|\mathcal{M}_a(\mathbb{F}_q)| = \sum_{\kappa\in\hat{I}_a^\Gamma} O_a^\kappa(1_D) \tag{8}$$

avec

$$O_a^\kappa(1_D) = \prod_{v\in|X|} \sum_{\substack{\gamma_v\in\mathfrak{g}(F_v)/\sim\\ \chi(\gamma_v)=a}} \langle \mathrm{inv}_v(\gamma_0, \gamma_v), \kappa\rangle\, O_{\gamma_v}(1_{D_v}) \tag{9}$$

qui ne dépend pas du choix de $\gamma_0 \in \mathfrak{g}(F)$. Nous appelons cette expression la κ-décomposition de (3). Elle ne dépend en fait pas du choix de l'élément γ_0. Ceci suggère l'existence d'une décomposition naturelle de la cohomologie de la fibre $\mathcal{M}_a$.

D'après [15] et [12, 9.6], cette κ-décomposition s'organise quand a varie en une somme en les classes $[\kappa]$ de $\hat{G}$-conjugaison des éléments d'ordre fini de $\hat{G}$. À la suite de (8), la partie anisotropique de (2) se récrit comme suit

$$|\mathcal{M}^{\mathrm{ani}}(\mathbb{F}_q)| = \sum_{[\kappa]\in\hat{G}/\sim} \sum_{a\in\mathcal{A}^{\mathrm{ani}}(\mathbb{F}_q)} \sum_{\kappa\in\hat{I}_a^\Gamma\cap[\kappa]} O_a^\kappa(1_D). \tag{10}$$

Précisons ce que nous entendons par $\kappa \in \hat{I}_a^\Gamma \cap [\kappa]$. Il existe un plongement du tore dual $\hat{I}_a$ du centralisateur I_a dans $\hat{G}$ bien défini à $\hat{G}$-conjugaison près. Le sous-ensemble de $\hat{I}_a^\Gamma$ des éléments dont l'image dans $\hat{G}$ appartient à la classe de conjugaison $[\kappa]$ est donc bien défini. Cette formule ci-dessus suggère l'existence d'une $[\kappa]$-décomposition de la restriction de $m_*\mathbb{Q}_\ell$ à l'ouvert $\mathcal{A}^{\mathrm{ani}}$.

Nous allons maintenant analyser la condition nécessaire pour que l'intersection $\hat{I}_a \cap [\kappa]$ soit non vide. Pour simplifier, supposons que G est un groupe semi-simple adjoint déployé. La monodromie du tore I_a est donnée par un homomorphisme $\Gamma \to W$ d'image un sous-groupe Σ_a de W bien défini à W-conjugaison près. Soit κ un représentant de $[\kappa]$ appartenant au tore maximal $\hat{T}$ de $\hat{G}$. Soit W_κ le sous-groupe de W des éléments qui fixent κ. Pour que $\hat{I}_a \cap [\kappa]$ soit non vide, il est nécessaire que Σ_a soit conjugué à un sous-groupe de W_κ. Si on suppose en plus que Σ_a est conjugué à $[W_\kappa]$, alors le cardinal de l'intersection $\hat{I}_a \cap [\kappa]$ est égal à $|\mathrm{Nor}(W_\kappa)|/|W_\kappa|$.

4. Symétries de la fibration de Hitchin

Dans le cas $G = \mathrm{GL}(n)$, pour tout $a = (a_i) \in \mathcal{A}(\bar{k})$, on a une courbe spectrale Y_a définie comme la courbe d'équation

$$t^n - a_1 t^{n-1} + \cdots + (-1)^n a_n = 0$$

tracée sur l'espace total du fibré en droites $\mathcal{O}_X(D)$. Si $a \in \mathcal{A}^\heartsuit(\bar{k})$, la courbe spectrale Y_a est réduite et la fibre $\mathcal{M}_a$ est la jacobienne compactifiée de Y_a classifiant les $\mathcal{O}_{Y_a}$-modules sans torsion de rang générique 1. Le groupe de symétries naturelles de Y_a est dans ce cas la jacobienne de Y_a classifiant les $\mathcal{O}_{Y_a}$-modules inversibles.

La construction de ces symétries dans le cas général repose sur une propriété simple des centralisateurs. Soit I le $\mathfrak{g}$-schéma en groupes des centralisateurs

$$I_x = \{g \in G \mid gxg^{-1} = x\}.$$

Soit $\mathfrak{g}^{\mathrm{reg}}$ l'ouvert de $\mathfrak{g}$ des éléments réguliers de $\mathfrak{g}$. Soit $\chi : \mathfrak{g} \to \mathfrak{t}/W$ le morphisme caractéristique de Chevalley *cf.* (1). Rappelons qu'on a une action naturelle de $\mathbb{G}_m$ sur $\mathfrak{t}/W$ qui fait de χ un morphisme $\mathbb{G}_m$-équivariant. On renvoie à [21, 3.2] pour la démonstration du lemme suivant.

Lemme 4.1. *Il existe un unique schéma en groupes affine lisse $\mathbb{G}_m$-équivariant J sur $\mathfrak{t}/W$ muni d'un homomorphisme G-équivariant $\chi^* J \to I$ dont la restriction à $\mathfrak{g}^{\mathrm{reg}}$ est un isomorphisme.*

Un point $a \in \mathcal{A}^\heartsuit(\bar{k})$ définit un morphisme $a : X \to [(\mathfrak{t}/W)/\mathbb{G}_m]$. L'image inverse $J_a = a^*[J/\mathbb{G}_m]$ est un schéma en groupes lisse sur X qui sur un ouvert dense U_a est un tore. Considérons le champ P_a des J_a-torseurs. On a une action de P_a sur $\mathcal{M}_a$ qui résulte de l'homomorphisme $\chi^* J \to I$. De plus, lorsque a varie dans $\mathcal{A}^\heartsuit$, les P_a s'organisent en un champ de Picard relatif lisse [21]. Cette construction a été inspirée par la lecture de [6]. Le champ de Picard P_a a été considéré aussi dans [5].

D'après un théorème de Grothendieck [8, 15.6.4], il existe un ouvert de Zariski P^0 de P telle que la fibre P_a^0 est la composante neutre de P_a, car P est lisse sur $\mathcal{A}^\heartsuit$. En prenant le faisceau quotient P/P^0, on obtient un faisceau en groupes abéliens $\pi_0(P)$ pour la topologie étale de $\mathcal{A}^\heartsuit$ tel que pour tout $a \in \mathcal{A}^\heartsuit(\bar{k})$, la fibre $\pi_0(P)_a$ est le groupe des composantes connexes $\pi_0(P_a)$. On a une description très précise de ce faisceau dans le cas où G est un groupe semi-simple adjoint.

Soit $X \times \mathcal{A}^\heartsuit \to [(\mathfrak{t}/W)/\mathbb{G}_m]$ le morphisme tautologique. Soit U l'image inverse de l'ouvert régulier semi-simple de $\mathfrak{t}/W$. Soit $\widetilde{U}$ le revêtement fini étale galoisien de groupe de Galois W défini comme l'image inverse du revêtement $\mathfrak{t} \to \mathfrak{t}/W$. Puisqu'on n'est dans $\mathcal{A}^\heartsuit$, le morphisme $\widetilde{U} \to \mathcal{A}^\heartsuit$ est un morphisme lisse à fibres non vides. D'après [8, 15.6.4] et [21, 6.2], il existe un faisceau $\pi_0(\widetilde{U})$ pour la topologie étale de $\mathcal{A}^\heartsuit$ tel que pour tout $a \in \mathcal{A}^\heartsuit(\bar{k})$, la fibre de $\pi_0(\widetilde{U})$ est l'ensemble des composantes connexes de $\widetilde{U}_a$. Notons que W agit sur $\pi_0(\widetilde{U})$ et que cette action est transitive fibre par fibre.

Proposition 4.2. *Il existe un homomorphisme surjectif canonique*

$$\mathbb{X}^\vee \times^W \pi_0(\widetilde{U}) \to \pi_0(P) \tag{11}$$

qui est un isomorphisme si G est un groupe semi-simple adjoint. Ici $\mathbb{X}^\vee$ est le groupe des cocaractères du tore T muni de l'action de W et le signe $\times^W$ désigne un produit contracté par l'action diagonale de W.

Pour tout $a \in \mathcal{A}^\heartsuit(\bar{k})$, en vertu du lemme d'homotopie, le groupe P_a agit sur $\mathrm{H}^i(\mathcal{M}_a, \mathbb{Q}_\ell)$ à travers le groupe des composantes connexes $\pi_0(P_a)$. Sur l'ouvert anisotropique $\mathcal{A}^{\mathrm{ani}}$, les groupes $\pi_0(P_a)$ sont finis. On peut donc décomposer

$$\mathrm{H}^i(\mathcal{M}_a, \mathbb{Q}_\ell) = \bigoplus_\kappa \mathrm{H}^i(\mathcal{M}_a, \mathbb{Q}_\ell)_\kappa \tag{12}$$

où κ parcourt l'ensemble des caractères du groupe $\pi_0(P_a)$. Ce groupe est un quotient de $\mathbb{X}^\vee_\Gamma$ si bien que ses caractères sont des éléments de $\hat{T}^\Gamma$. Cette décomposition correspond via le dictionnaire faisceaux-fonctions à la formule (8). En effet, d'après [21, 4.6] le champ quotient $[\mathcal{M}_a/P_a]$, est un produit des champs définis localement de la même façon qu'une intégrale orbitale globale divisée par un nombre de Tamagawa du centralisateur est égal au produit des intégrales orbitales locales.

Globalement sur $\mathcal{A}^{\mathrm{ani}}$, le groupe P agit sur les faisceaux de cohomologie perverse ${}^p\mathrm{H}^i(m^{\mathrm{ani}}_* \mathbb{Q}_\ell)$ et cette action se factorise à travers le faisceau $\pi_0(P)$ en vertu d'une variante du lemme d'homotopie [20, 3.2.3]. Pour chaque ouvert étale U de $\mathcal{A}^{\mathrm{ani}}$, soit $\pi_0(P)(U)^*$ le groupe des caractères de $\pi_0(P)(U)$ et soit $\pi_0(P)^{\mathrm{co}}$ le cofaisceau associé à ce précofaisceau $U \mapsto \pi_0(P)^{\mathrm{co}}$. On envoie à [21, §8] pour un petit résumé de la notion de cofaisceau. On en déduit une décomposition de ${}^p\mathrm{H}^i(m^{\mathrm{ani}}_* \mathbb{Q}_\ell)$ selon l'ensemble des sections globales du cofaisceau $\pi_0(P)^{\mathrm{co}}$. La description (11) du faisceau $\pi_0(P)$ permet de définir une application canonique de l'ensemble des sections globales $\Gamma(\mathcal{A}^{\mathrm{ani}}, \pi_0(P)^{\mathrm{co}})$ dans l'ensembles des classes de $\hat{G}$-conjugaison $[\kappa]$ des éléments κ d'ordre fini de $\hat{G}$, voir [21, 8.4]. On en déduit une décomposition

$${}^p\mathrm{H}^i(m^{\mathrm{ani}}_* \mathbb{Q}_\ell) = \bigoplus_{[\kappa]} {}^p\mathrm{H}^i(m^{\mathrm{ani}}_* \mathbb{Q}_\ell)_{[\kappa]} \tag{13}$$

qui correspond à (10) au niveau des fonctions.

Pour simplifier les notations, considérons le faisceau pervers gradué

$$(m^{\mathrm{ani}}_{*,\mathrm{gr}} \mathbb{Q}_\ell)_{[\kappa]} = \bigoplus_{i \in \mathbb{Z}} {}^p\mathrm{H}^i(m^{\mathrm{ani}}_* \mathbb{Q}_\ell)_{[\kappa]}.$$

On peut estimer le support de $(m^{\mathrm{ani}}_{*,\mathrm{gr}} \mathbb{Q}_\ell)_{[\kappa]}$. Pour simplifier, supposons de nouveau G semi-simple adjoint et déployé. On peut stratifier $\mathcal{A}$ de la façon suivante

$$\mathcal{A} = \bigsqcup_{[\Sigma]} \mathcal{A}_{[\Sigma]}$$

où $[\Sigma]$ parcourt l'ensemble des classes de conjugaison de sous-groupes Σ de W. Un point géométrique $a \in \mathcal{A}(\bar{k})$ appartient à la strate $[\Sigma]$ si le groupe de monodromie du tore I_a est $[\Sigma]$. Une autre strate $\mathcal{A}_{[\Sigma']}$ est incluse dans l'adhérence de $\mathcal{A}_{[\Sigma]}$ si Σ' est conjuguée à un sous-groupe de Σ.

Proposition 4.3. *Le morceau ${}^p\mathrm{H}^i(m_*^{\mathrm{ani}}\mathbb{Q}_\ell)_{[\kappa]}$ est supporté par $\mathcal{A}_{[W_\kappa]}$ où W_κ est le fixateur dans W d'un représentant $\kappa \in \hat{T}$ de $[\kappa]$.*

La cohomologie perverse de la fibration de Hitchin se décompose donc en morceaux de différents supports. On appelle stables les morceaux correspondant aux $\kappa \in Z_{\hat{G}}$. Son support est a priori tout l'espace $\mathcal{A}^{\mathrm{ani}}$. Il y a lieu de penser que ces morceaux stables ne contiennent pas de facteurs directs de support strictement plus petit. Ceci est probablement une conjecture difficile.

5. Groupes endoscopiques

Pour simplifier l'exposition, supposons que G est un groupe semi-simple adjoint déployé. Le groupe dual $\hat{G}$ est alors un groupe semi-simple simplement connexe muni de l'action triviale de Γ. Soit $\kappa \in \hat{G}$ un élément d'ordre fini. Le centralisateur $\hat{G}_\kappa$ est alors un groupe réductif connexe qu'on notera $\hat{H}$. Soit H le groupe déployé sur k dont le dual est $\hat{H}$. Il n'y a qu'une faible relation entre H et G : ils partagent un tore maximal T et le groupe de Weyl W_H de H est un sous-groupe de réflexions du groupe de Weyl W de G. Avec l'hypothèse G adjoint, W_H est le fixateur W_κ de κ dans W.

Considérons l'espace de module des Hitchin pour le triplet (X, D, H) qui paramètre les couples (E_H, ϕ) où E_H est un H-torseurs sur X et où ϕ est une section globale de $\mathrm{ad}(E_H)(D)$. On a aussi une fibration de Hitchin

$$n : \mathcal{N} \to \mathcal{B}$$

où $\mathcal{B}$ est l'espace des sections globales du fibré $(\mathfrak{t}/W_H) \times^{\mathbb{G}_m} L_D$. Le morphisme évident $\mathfrak{t}/W_H \to \mathfrak{t}/W$ induit un morphisme $\pi : \mathcal{B} \to \mathcal{A}$. Par restriction à $\mathcal{B}^\heartsuit$, on obtient un morphisme non ramifié $\pi^\heartsuit : \mathcal{B}^\heartsuit \to \mathcal{A}^\heartsuit$ d'image $\bar{\mathcal{A}}_{W_H}$. Au-dessus de l'ouvert $\mathcal{A}_{W_H}$ de $\bar{\mathcal{A}}_{W_H}$, le morphisme

$$\pi_{W_H} : \mathcal{B}_{W_H} \to \mathcal{A}_{W_H}$$

est un morphisme fini et étale de degré $|\mathrm{Nor}(W_H)|/|W_H|$ égal au cardinal de l'ensemble $\hat{I}_a^\Gamma \cap [\kappa]$ pour tout point géométrique $a \in \mathcal{A}_{W_H}(\bar{k})$.

Sur $\mathcal{B}^\heartsuit$, on a un champ de Picard Q des symétries de la fibration de Hitchin $n : \mathcal{N} \to \mathcal{B}$. L'action qui s'en déduit sur $n_*^\heartsuit\mathbb{Q}_\ell$ se factorise à travers le faisceau $\pi_0(Q)$. Sur la strate ouverte $\mathcal{N}_{\Sigma_H}$ de $\mathcal{N}$, le faisceau $\pi_0(Q)$ est le faisceau constant de valeur $Z_{\hat{H}}$.

Dans la décomposition $n_*^{\mathrm{ani}}\mathbb{Q}_\ell$, on a donc un morceau stable $(n_*^{\mathrm{ani}}\mathbb{Q}_\ell)_\kappa$ correspondant à l'élément $\kappa \in [\kappa]$ qui a servi à définir $\hat{H}$. De même, on a un κ-morceau

$(\pi_{W_H}^* m_*^{\mathrm{ani}}\mathbb{Q}_\ell)_\kappa$. D'après la proposition 4.3, on sait que ces deux morceaux sont sommes directes de faisceaux pervers purs. Suivant Langlands et Shelstad [16], il serait tentant de formuler la conjecture suivante.

Conjecture 5.1. Il existe
- un nombre entier naturel d
- un système local $\mathcal{L}$ de rang 1 et d'ordre 2 sur $\mathcal{B}_{\Sigma_H}$
- un $Z_{\hat{H}}$-torseur Υ sur $\mathcal{B}_{\Sigma_H}$

tels qu'il existe un isomorphisme de faisceaux pervers gradués au-dessus de $\mathcal{B}_{W_H}$

$$((n_*^{\mathrm{ani}}\mathbb{Q}_\ell)_\kappa^\Upsilon \otimes \mathcal{L}[-2d](-d)) \to (\pi_{W_H}^* m_*^{\mathrm{ani}}\mathbb{Q}_\ell)_\kappa$$

où l'exposant Υ consiste à tordre l'objet muni d'une action de $Z_{\hat{H}}$ par le $Z_{\hat{H}}$-torseur Υ, où $[-2d]$ est le décalage dans la graduation et où $(-d)$ est le twist à la Tate.

Par adjonction, on déduit de cette conjecture un isomorphisme de faisceaux pervers gradués sur $\mathcal{A}_{W_H}$

$$(\pi_{W_H})_*((n_*^{\mathrm{ani}}\mathbb{Q}_\ell)_\kappa^\Upsilon \otimes \mathcal{L}[-2d](-d)) \to (m_*^{\mathrm{ani}}\mathbb{Q}_\ell)_{[\kappa]}.$$

Il est très tentant de penser que cet isomorphisme s'étend à $\bar{\mathcal{A}}_{W_H}$ car en mettant ensemble ces isomorphismes pour les différents groupes endoscopiques, on obtiendrait géométriquement la stabilisation complète de la partie anisotrope de la formule des traces. Pour le moment, il manque un peu d'exemples pour que cette conjecture plus générale soit complètement convaincante.

Soit $b \in \mathcal{B}_{W_H}(\mathbb{F}_q)$ un $\mathbb{F}_q$-point de $\mathcal{B}_{W_H}$ d'image $a \in \mathcal{A}_{W_H}(\mathbb{F}_q)$. La conjecture 5.1 implique l'égalité suivante qui est une forme globale du lemme fondamental

$$O_a^\kappa(1_D) = q^d \epsilon_b(\mathcal{L})\epsilon_b(\kappa, \Upsilon) O_{H,b}^\kappa(1_D).$$

Ici $\epsilon_b(\mathcal{L}) \in \{\pm 1\}$ est le signe défini par la fibre en b du système local $\mathcal{L}$ de rang 1 d'ordre 2 et $\epsilon_b(\kappa, \Upsilon)$ est la racine d'unité définie par la fibre en b du système local de rang 1 obtenu en poussant le $Z_{\hat{H}}$-torseur Υ par le caractère κ. Les intégrales $O_a^\kappa(1_D)$ et $O_{H,b}^\kappa(1_D)$ sont des produits de κ-intégrales orbitales locales. La κ-intégrale orbitale pour H est une intégrale orbitale stable car $\kappa \in Z_{\hat{H}}$. Il est très probable qu'on peut déduire la forme usuelle locale du lemme fondamental, conjecturée par Langlands et Shelstad [16], à partir de cette forme globale. En particulier, le nombre $q^d \epsilon_b(\mathcal{L})\epsilon_b(\kappa, \Upsilon)$ doit être le produit en toutes les places de facteurs de transfert de Langlands–Shelstad.

Pour formuler la conjecture 5.1 de façon plus précise, il est nécessaire de construire les différents ingrédients d, $\mathcal{L}$ et Υ. Dans le cas du groupe unitaire, d et $\mathcal{L}$ ont été construits dans [20], et Υ est trivial. Les constructions de d et $\mathcal{L}$ se généralisent sans trop de difficultés au cas général. Il existe aussi en général un candidat naturel pour Υ dont la construction reste toutefois conditionnelle.

Dans le cas des groupes unitaires, la conjecture a été démontrée dans [20]. Le cas général reste ouvert.

Remerciements. J'exprime ma reconnaissance à J.-F. Dat, V. Drinfeld, A. Genestier, R. Kottwitz, J.-P. Labesse, L. Lafforgue, R. Langlands, C. Moeglin, M. Rapoport et tout particulièrement à G. Laumon pour l'encouragement et l'aide qu'ils m'ont apportés lors de différents stades de ce projet. Je remercie G. Laumon et Ngô Dac Tuân pour leur relecture attentive de ce rapport.

Références

[1] Arthur, J., Toward a stable trace formula. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 507–517.

[2] Biswas, I., Ramanan, S., Infinitesimal study of Hitchin pairs. *J. London Math. Soc.* **49** (1994), 219–231.

[3] Cluckers, R., Loeser, F., Constructible exponential functions, motivic Fourier transform and transfer principle. Prépublication.

[4] Dat, J.-F., Lemme fondamental et endoscopie, une approche géométrique. *Séminaire Bourbaki* n° **940**, novembre 2004.

[5] Donagi, R., Gaitsgory, D., The gerbs of Higgs bundles. *Transform. Groups* **7** (2002), 109–153.

[6] Faltings, G., Stable G-bundles and projective connections. *J. Alg. Geom.* **2** (1993), 507–568.

[7] Goresky, M., Kottwitz, R., MacPherson, R., Homology of affine Springer fiber in the unramified case. *Duke Math. J.* **121** (2004), 509–561.

[8] Grothendieck, A., Dieudonné J., Éléments de géométrie algébrique IV.3. *Inst. Hautes Études Sci. Publ. Math.* **28** (1966), 5–255.

[9] Hales, T., The fundamental lemma for Sp(4). *Proc. Amer. Math. Soc.* **125** (1) (1997), 301–308.

[10] Hitchin, N., Stable bundles and integrable connections. *Duke Math. J.* **54** (1987), 91–114.

[11] Kottwitz, R., Harmonic Analysis on semi-simple p-adic Lie algebras. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 553–562.

[12] Kottwitz, R., Stable trace formula : elliptic singular terms. *Math. Ann.* **275** (1986), 365–399.

[13] Jacquet, H., and Langlands, R., *Automorphic forms on* GL(2). Lecture Notes in Math. 114, Springer-Verlag, Berlin, New York 1970.

[14] Labesse, J.-P., and Langlands, R., L-indistinguishability for SL(2). *Canad. J. Math.* **31** (1979), 726–785.

[15] Langlands, R., *Les débuts d'une formule des traces stables*. Publications de l'Université Paris VII, vol. 13, 1983.

[16] Langlands, R., and Shelstad, D., On the definition of transfer factors. *Math. Ann.* **278** (1987), 219–271.

[17] Laumon, G., Fibres de Springer et jacobiennes compactifiées. Prépublication ; arXiv math.AG/0204109.

[18] Laumon, G., Sur le lemme fondamental pour les groupes unitaires. Prépublication ; arXiv math.AG/0212245.

[19] Laumon, G., et Moret-Bailly, L., *Champs algébriques*. Ergeb. Math. Grenzgeb. (3) 39, Springer-Verlag, Berlin 2000.

[20] Laumon, G., et Ngô, B.-C., Le lemme fondamental pour les groupes unitaires. Prépublication ; arXiv math.AG/0404454.

[21] Ngô, B.-C., Fibration de Hitchin et endoscopie. *Invent. Math.*, à paraître.

[22] Ono, T., On Tamagawa numbers. In *Algebraic groups and discontinuous subgroups* (Boulder, Colo., 1965), Proc. Sympos. Pure Math. 9, Amer. Math. Soc., Providence, RI, 1966, 122–132.

[23] Rogawski, J., *Automorphic representations of unitary groups in three variables.* Ann. of Math. Stud. 123, Princeton University Press, Princeton, NJ, 1990.

[24] Waldspurger, J.-L., Sur les intégrales orbitales tordues pour les groupes linéaires : un lemme fondamental. *Canad. J. Math.* **43** (1991), 852–896.

[25] Waldspurger, J.-L., Comparaison d'intégrales orbitales pour des groupes p-adiques. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 807–816.

[26] Waldspurger, J.-L., Le lemme fondamental implique le transfert. *Compositio Math.* **105** (1997), 153–236.

[27] Waldspurger, J.-L., Endoscopie et changement de caractéristique. Prépublication.

[28] Weissauer, R., A special case of fundamental lemma. Prépublication.

Département de mathématiques, Université de Paris-Sud, 91405 Orsay, France
E-mail: Bao-Chau.Ngo@math.u-psud.fr

Hecke algebras and harmonic analysis

Eric M. Opdam*

Abstract. Iwahori–Hecke algebras are ubiquitous. One encounters these algebras in subjects as diverse as harmonic analysis, equivariant K-theory, orthogonal polynomials, quantum groups, knot theory, algebraic combinatorics, and integrable models in statistical physics. In this exposition we will mostly concentrate on the analytic aspects of affine Hecke algebras and study them from the perspective of operator algebras. We will discuss the Plancherel theorem for these type of algebras, and based on a conjectural invariance property of their (operator algebraic) K-theory, study the structure of the tempered dual.

Mathematics Subject Classification (2000). Primary 20C08; Secondary 22D25, 43A90.

Keywords. Hecke algebra, Plancherel measure, tempered dual, K-theory.

1. Introduction

Let G be a group and let $K \subset G$ be an almost normal subgroup of G, i.e. a subgroup whose double cosets are finite unions of one-sided cosets. The Hecke algebra of the pair (G, K) is the convolution algebra of $\mathbb{Z}$-valued functions with finite support on the double coset space $K\backslash G/K$. More generally, given a K-module (M, σ) (over some commutative, unital ring R) one considers the convolution algebra $H_R(G, K, \sigma)$ of $\mathrm{End}_R(M^\vee)$-valued $(K, \sigma^\vee)$-spherical functions which are supported on finitely many double cosets of K. Hecke algebras were introduced in this abstract setting by Shimura in the 1950s, following the original work of Hecke on certain linear operators acting in a space of modular forms. The study of representations of Hecke algebras in spaces of modular forms is of basic importance for the study of modular forms.

Later it became apparent that this concept is also fundamental for understanding the representation theory of finite reductive groups, and this seems also true for p-adic reductive groups. Let k be a non-archimedean local field and let G be the group of k-points of a connected reductive group defined over k, equipped with the locally compact, totally disconnected Hausdorff topology it inherits from k. Any compact open subgroup $K \subset G$ is almost normal, hence in this situation we have a large supply of Hecke algebras of the form $H_{\mathbb{C}}(G, K, \sigma)$ where (V_σ, σ) is a complex finite dimensional smooth representation of K.

*The author is grateful to the Netherlands Organization for Scientific Research (NWO) for supporting this research by means of a Pionier grant.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

The fundamental and beautiful result which is the underpinning of the application of these Hecke algebras to the representation theory of G is Bernstein's Decomposition Theorem [11]. It states that the category of smooth representations of G has a canonical decomposition as a product of blocks $\mathfrak{R}_{\mathfrak{s}}$ which are parametrized by the components $\mathfrak{s}$ of the Bernstein variety Σ of "supercuspidal pairs" (L, ρ) modulo G-conjugacy, where $L \subset G$ is a Levi-subgroup, and where ρ is an irreducible supercuspidal representation of L, i.e. a representation whose matrix coefficients are compactly supported modulo the center of L. Each block is by construction equivalent to the category of modules over a two sided ideal $C_c^\infty(G)_{\mathfrak{s}}$ of the convolution algebra $C_c^\infty(G)$ of compactly supported, locally constant complex valued functions on G (for the sake of this exposition I will refrain from calling $C_c^\infty(G)$ the Hecke algebra of G, although it is customary to so).

It is a major question how to describe the blocks $\mathfrak{R}_{\mathfrak{s}}$. Bushnell and Kutzko [16], [17] introduced the notion of $\mathfrak{s}$-types. A pair (K, σ) is called an $\mathfrak{s}$-type if the block $\mathfrak{R}_{\mathfrak{s}}$ consists precisely of those smooth representations of G which are generated by their (K, σ) isotypic component. In that case the functor $V \to \mathrm{Hom}_K(V_\sigma, V|_K)$ is an equivalence from $\mathfrak{R}_{\mathfrak{s}}$ to the category of $\mathcal{H}_{\mathbb{C}}(G, K, \sigma)$-modules. This notion of types originates from the work of Borel [14], who showed that an irreducible smooth G-module V is a subquotient of the unramified principal series iff V contains fixed vectors with respect to an Iwahori subgroup $B \subset G$ (the corresponding component of Σ is called the "Borel component"). Through the work of Bushnell and Kutzko (loc. cit.), Morris [51], [52], and Moy and Prasad [53] $\mathfrak{s}$-types are known to exist in many cases. For instance $\mathfrak{s}$-types always exist for "level 0" components $\mathfrak{s}$.

The algebra $H_{\mathbb{C}}(G, K, \sigma)$ comes equipped with a trace $\mathrm{tr}\colon f \to \mathrm{trace}(f(e))$ and a $*$ structure $f^*(g) = f(g^{-1})^*$. This defines a canonical C^*-algebra closure $C_r^*(H, K, \sigma)$ of $H_{\mathbb{C}}(G, K, \sigma)$ (the reduced C^*-algebra) which is of type I and comes with the distinguished faithful trace tr. The equivalence between a block $\mathfrak{R}_{\mathfrak{s}}$ and the module category of the Hecke algebra $H_{\mathbb{C}}(G, K, \sigma)$ of an $\mathfrak{s}$ respects this structure, and the Plancherel measure of G restricted to the irreducible representations in the block $\mathfrak{R}_{\mathfrak{s}}$ coincides with the spectral measure of the trace tr. We normalize tr to obtain a tracial state τ on $C_r^*(H, K, \sigma)$, and we refer to its spectral measure as the "Plancherel measure of $H_{\mathbb{C}}(G, K, \sigma)$". The theory of types seeks to decompose the harmonic analysis on G essentially in two separate parts: (1) knowledge of the supercuspidal representations of all Levi subgroups $L \subset G$ and (2) knowledge of the Plancherel measure of the Hecke algebras $H_{\mathbb{C}}(G, K, \sigma)$ [15].

As a complement to these results, the structure of the Hecke algebra $H_{\mathbb{C}}(G, K, \sigma)$ of an $\mathfrak{s}$-type can be described fairly explicitly in various cases. In the case of the Borel component this algebra is the *Iwahori–Hecke algebra* $\mathcal{H}(W, q)$ [32], where W is an extended affine Weyl group which can be attached to G by Bruhat–Tits theory, and where q is a label function on W which is defined in terms of the structure of a (generalized) affine BN-pair for G and the cardinality $\boldsymbol{q}$ of the residue field of k. By results of Morris (loc. cit.) and Lusztig [41] the Hecke algebra $H_{\mathbb{C}}(G, K, \sigma)$ of a type (K, σ) of level 0 is always a twisted crossed product of an Iwahori–Hecke algebra of

the form $\mathcal{H}(W', q')$ (for a certain affine Weyl group W' and label function q') and a group $C(K, \sigma)$ (where the 2-cocycle lives on $C(K, \sigma)$). These results draw heavily on the work of Howlett and Lehrer [31] who successfully followed a similar approach for the representation theory of finite groups of Lie type.

The above exposition makes a quite compelling case for the study of an Iwahori–Hecke algebra $\mathcal{H}(W, q)$ as an object of (harmonic) analysis and the spectral problem described above. Indeed, this point of view did not go unnoticed and in some sense was already promoted by Matsumoto in [47]. But it turns out that it is quite difficult to carry it out. The description of the support of the Plancherel measure amounts to the description of the tempered dual of $\mathcal{H}(W, q)$. Using geometric methods of a completely different nature this problem was solved explicitly by Kazhdan and Lusztig in their profound paper [34], in the special case of the "Borel component" when in addition G is split semisimple and of adjoint type. Lusztig [41], [42] has in principle solved such classification problems in greater generality, when G splits over an unramified extension of k, and σ is a cuspidal unipotent representation. These methods do not give information on the Plancherel measures.

Iwahori–Hecke algebras also play a fundamental role in a wide range of other areas for some of which the aforementioned spectral problems are of immediate interest, such as integrable models in mathematical physics (the Calogero–Moser systems [29], [54], and also the generalized quantum Bose gas with delta function potential and the nonlinear Schrödinger equation [27], [24]), and the theory of multivariable orthogonal polynomials and special functions [46], [19]. These applications have led to interesting new directions in the theory of Hecke algebras (most notably Ivan Cherednik's double affine Hecke algebra [18], [19]) and this in fact raises challenging new questions in harmonic analysis. We also mention the role of the Hecke algebra for unitarizability of Iwahori-spherical representations [5], [6].

Therefore it is a problem of considerable interest to describe the Plancherel measure of the Iwahori–Hecke algebras $\mathcal{H}(W, q)$ of affine type (simply called "affine Hecke algebras" in the sequel) explicitly, and it is this problem that we will address in this paper. The paper has three parts. In the sections 2–4 we review results of [56] on the L^2-completion of the affine Hecke algebra. The main results are: (1) An algebraic characterization of the central support of the tempered spectrum. (2) The Plancherel density depends up to constants independent of $\boldsymbol{q}$ only on the central character. (3) An explicit product formula for the formal dimensions of the discrete series, up to constants independent of $\boldsymbol{q}$. The sections 5–6 give an overview of the joint work of Patrick Delorme and myself [22], [23] on the Schwartz algebra completion of the affine Hecke algebra. Here we discuss the geometric structure of the tempered dual by means of the analogue of results of Harish-Chandra [25], [26] and Knapp–Stein [36] on analytic R-groups. Finally in Sections 7–8 we discuss various natural conjectures on the K-theory of the Schwartz algebra. In the three parameter example of type C_n^{aff} we indicate how these conjectures lead in fact to complete description of the tempered dual. In striking contrast to the geometric methods mentioned above, the affine Hecke algebras with generic unequal parameters should be considered as

the most basic cases from this point of view. Using the conjectures, all non-generic cases are understood by deformation to the generic case.

2. Affine Hecke algebras

The structure of an affine Hecke algebra $\mathcal{H} = \mathcal{H}(\mathcal{R}, q)$ is determined by an affine root datum (with basis) $\mathcal{R}$ together with a label function q defined on the extended affine Weyl group W associated to $\mathcal{R}$. We refer the reader to [39], [56], [22] for the details of the definition of the algebra $\mathcal{H}(\mathcal{R}, q)$, which we will only briefly review here.

Let $\mathcal{R} = (X, R_0, Y, R_0^\vee, F_0)$ be a root datum (with basis $F_0 \subset X$ of simple roots of $R_0 \subset X$). This means that R_0 is a (reduced, integral) root system with basis of simple roots F_0, that $R_0^\vee$ is the coroot system of R_0, and that X, Y are lattices in duality such that $R_0 \subset X$ and $R_0^\vee \subset Y$. For example take $X = P(R_0)$, the weight lattice of R_0, and $Y = Q(R_0^\vee)$, the root lattice of $R_0^\vee$. If $\mathfrak{a} := \mathbb{R} \otimes_{\mathbb{Z}} Y$ is spanned by the coroots we call $\mathcal{R}$ semisimple.

Let $W_0 = W(R_0)$ denote the Weyl group of the reduced integral root system R_0. The extended affine Weyl group W associated with $\mathcal{R}$ is by definition $W = W_0 \ltimes X$. The affine root system R is equal to $R := R_0^\vee \times \mathbb{Z} \subset Y \times \mathbb{Z}$. We view elements of $Y \times \mathbb{Z}$ as affine linear functions on X with values in $\mathbb{Z}$. Observe that R is closed for the natural action of W on the set of integral affine linear functions $Y \times \mathbb{Z}$ on X. Furthermore R is the disjoint union of the sets of positive and negative affine roots $R = R_+ \cup R_-$ as usual, and we define the length function l on W by

$$l(w) := |R_+ \cap w^{-1} R_-|. \tag{2.1}$$

A label function $q\colon W \to \mathbb{C}^\times$ is a function which is length multiplicative (i.e. $q(uv) = q(u)q(v)$ if $l(uv) = l(u) + l(v)$) and which in addition satisfies $q(\omega) = 1$ if $l(\omega) = 0$. Thus a label function is completely determined by its values on the set S^{aff} of affine simple reflections in W. It follows easily that its restriction to S^{aff} is constant on W-conjugacy classes of simple reflections. Conversely, any $\mathbb{C}^\times$-valued function on S^{aff} with this property extends uniquely to a label function.

For the purpose of this analytic approach to affine Hecke algebras we will work with positive real label functions only.

Definition 2.1. We denote the set of all positive real label functions for $\mathcal{R}$ by $\mathcal{Q} = \mathcal{Q}_{\mathcal{R}}$. For later reference, we choose a base $\boldsymbol{q} > 1$ and define $f_s \in \mathbb{R}$ such that $q(s) = \boldsymbol{q}^{f_s}$ for all $s \in S^{\mathrm{aff}}$.

Definition 2.2. Given a root datum $\mathcal{R}$ and a positive real label function $q \in \mathcal{Q}$ there exists a unique complex associative unital algebra $\mathcal{H}$ with $\mathbb{C}$-basis N_w ($w \in W$) subject to the following relations (here $q(s)^{1/2}$ denotes the positive square root of $q(s)$):

(a) $N_{uv} = N_u N_v$ for all $u, v \in W$ such that $l(uv) = l(u) + l(v)$.

(b) $(N_s + q(s)^{-1/2})(N_s - q(s)^{1/2}) = 0$ for all $s \in S^{\mathrm{aff}}$.

We call $\mathcal{H} = \mathcal{H}(\mathcal{R}, q)$ the affine Hecke algebra associated with the pair $(\mathcal{R}, q)$.

Remark 2.3. We equip $\mathcal{Q}$ in the obvious way with the structure of the vector group $\mathbb{R}_+^N$ where N denotes the number of W-conjugacy classes in S^{aff}. Given the base $\boldsymbol{q} > 1$ we identify $\mathcal{Q}$ with the finite dimensional real vector space of real functions $s \to f_s$ on S^{aff} which are constant on W-conjugacy classes (see Definition 2.1). In this sense we speak of (linear) *hyperplanes* in $\mathcal{Q}$ (this notion is independent of $\boldsymbol{q}$). By a *half line* in $\mathcal{Q}$ we mean a family of label functions $q \in \mathcal{Q}$ in which the $f_s \in \mathbb{R}$ are kept fixed and are not all equal to 0 and $\boldsymbol{q}$ is varying in $\mathbb{R}_{>1}$. As we will see later, for many problems it is interesting to consider the family of Hecke algebras when q varies in a half line in $\mathcal{Q}$ ("changing the base").

2.1. Root labels for the non-reduced root system. The label function q on W can also be defined in terms of root labels for a certain possibly non-reduced root system which is associated with $\mathcal{R}$. We define R_{nr} associated with $\mathcal{R}$ by

$$R_{\mathrm{nr}} := R_0 \cup \{2\alpha \mid \alpha^\vee \in R_0^\vee \cap 2Y\}. \tag{2.2}$$

Observe that $a + 2 \in Wa$ for all $a \in R$, but that $a + 1 \in Wa$ iff $a = \alpha^\vee + n$ with $2\alpha \notin R_{\mathrm{nr}}$. For affine simple roots $a \in F^{\mathrm{aff}}$ (and thus in particular for $a \in F_0^\vee$) we define

$$q_{a+1} := q(s_a), \tag{2.3}$$

and we extend this to a W-invariant function $a \to q_a$ on the affine root system R (this is possible in a unique fashion). Now for $\alpha = 2\beta \in R_{\mathrm{nr}} \backslash R_0$ we define

$$q_{\alpha^\vee} := \frac{q_{\beta^\vee + 1}}{q_{\beta^\vee}}. \tag{2.4}$$

In this way the set of label functions q on W corresponds bijectively to the set of positive W_0-invariant functions $R_{\mathrm{nr}} \ni \alpha \to q_{\alpha^\vee}$.

2.2. Bernstein presentation. There is another, extremely important presentation of the algebra $\mathcal{H}$, due to J. Bernstein (unpublished) and Lusztig [39]). Since the length function is additive on the dominant cone X^+, the map $X^+ \ni x \to N_x$ is a homomorphism of the commutative monoid X^+ with values in $\mathcal{H}^\times$, the group of invertible elements of $\mathcal{H}$. Thus there exists a unique extension to a homomorphism $X \ni x \to \theta_x \in \mathcal{H}^\times$ of the lattice X with values in $\mathcal{H}^\times$.

The abelian subalgebra of $\mathcal{H}$ generated by $\theta_x, x \in X$, is denoted by $\mathcal{A}$. Let $\mathcal{H}_0 = \mathcal{H}(W_0, q_0)$ be the finite type Hecke algebra associated with W_0 and the restriction q_0 of q to W_0. Then the Bernstein presentation asserts that both the collections $\theta_x N_w$ and $N_w \theta_x$ ($w \in W_0, x \in X$) are bases of $\mathcal{H}$ over $\mathbb{C}$, subject only to the cross relation

(for all $x \in X$ and $s = s_\alpha$ with $\alpha \in F_0$):

$$\theta_x N_s - N_s \theta_{s(x)} = \begin{cases} (q_{\alpha^\vee}^{1/2} - q_{\alpha^\vee}^{-1/2}) \frac{\theta_x - \theta_{s(x)}}{1-\theta_{-\alpha}} & \text{if } 2\alpha \notin R_{\mathrm{nr}}, \\ ((q_{\alpha^\vee/2}^{1/2} q_{\alpha^\vee}^{1/2} - q_{\alpha^\vee/2}^{-1/2} q_{\alpha^\vee}^{-1/2}) + (q_{\alpha^\vee}^{1/2} - q_{\alpha^\vee}^{-1/2})\theta_{-\alpha}) \frac{\theta_x - \theta_{s(x)}}{1-\theta_{-2\alpha}} & \text{if } 2\alpha \in R_{\mathrm{nr}}. \end{cases} \quad (2.5)$$

2.3. The center $\mathcal{Z}$ of $\mathcal{H}$. From the Bernstein presentation of $\mathcal{H}$ one easily derives the following fundamental result, the description of the center of $\mathcal{H}$.

Theorem 2.4 (Bernstein). *The center $\mathcal{Z}$ of $\mathcal{H}$ is equal to $\mathcal{A}^{W_0}$. In particular, $\mathcal{H}$ is finitely generated over its center.*

As an immediate consequence we see that irreducible representations of $\mathcal{H}$ are finite dimensional by application of (Dixmier's version of) Schur's lemma.

We denote by T the complex algebraic torus $T = \mathrm{Hom}(X, \mathbb{C}^\times)$ of complex characters of the lattice X. The space $\mathrm{Spec}(\mathcal{Z})$ of complex homomorphisms of $\mathcal{Z}$ is thus canonically isomorphic to the (categorical) quotient $W_0\backslash T$. By Bernstein's theorem and Schur's lemma we obtain a continuous, finite, surjective map

$$z\colon \mathrm{Irr}(\mathcal{H}) \to \mathrm{MaxSpec}(\mathcal{Z}) = W_0\backslash T, \quad [\pi] \mapsto z(\pi), \quad (2.6)$$

where $\mathrm{Irr}(\mathcal{H})$, the set of equivalence classes of irreducible representations of $\mathcal{H}$, is given the usual Jacobson topology via its identification with the primitive ideal spectrum of $\mathcal{H}$. We call this map the (algebraic) *central character*.

3. L^2-theory and abstract Plancherel theorem

We will study $\mathcal{H}$ via certain topological completions of $\mathcal{H}$. In this section we will study the L^2-completion of $\mathcal{H}$ and the associated reduced C^*-algebra of $\mathcal{H}$.

3.1. $\mathcal{H}$ as a Hilbert algebra. It is a basic fact that the anti-linear map $*$ on $\mathcal{H}$ defined by

$$\Big(\sum_{w\in W} c_w N_w\Big)^* = \sum_{w\in W} \overline{c_w} N_{w^{-1}} \quad (3.1)$$

is an anti-involution of $\mathcal{H}$, making $(\mathcal{H}, *)$ into an involutive algebra. In addition, the linear functional τ defined by

$$\tau\Big(\sum_{w\in W} c_w N_w\Big) = c_e \quad (3.2)$$

is a positive trace on $(\mathcal{H}, *)$. In particular, the sesquilinear pairing $(x, y) := \tau(x^* y)$ defines a pre-Hilbert structure on $\mathcal{H}$.

Definition 3.1. We call $L^2(\mathcal{H})$ the Hilbert space completion of $\mathcal{H}$. Observe that the elements N_w ($w \in W$) form a Hilbert basis for $L^2(\mathcal{H})$.

It is easy to see that the regular representation of $\mathcal{H}$ extends to a representation of $\mathcal{H}$ in $B(L^2(\mathcal{H}))$, the algebra of bounded linear operators on $L^2(\mathcal{H})$. This gives $\mathcal{H}$ the structure of a unital Hilbert algebra, with its Hermitian form defined by the finite positive trace τ.

3.2. The reduced C^*-algebra $\mathfrak{C}$ of $\mathcal{H}$. The following results on the reduced C^*-algebra of $\mathcal{H}$ go back to [47].

Definition 3.2. We define the reduced C^*-algebra $\mathfrak{C}$ of $\mathcal{H}$ as the norm closure of $\lambda(\mathcal{H}) \subset B(L^2(\mathcal{H}))$, where λ denotes the left regular representation of $\mathcal{H}$. We identify $\mathfrak{C}$ with a dense subspace of $L^2(\mathcal{H})$ via the continuous injection $\mathfrak{C} \ni x \to x(1) \in L^2(\mathcal{H})$.

Let λ (resp. ρ) denote the left (resp. right) regular representation of $\mathfrak{C}$ on $L^2(\mathcal{H})$. One has the following basic statements:

Corollary 3.3. *The C^*-algebra completion $\mathfrak{C}$ of $\mathcal{H}$ has type I, and τ extends to a finite tracial state of $\mathfrak{C}$ such that $\lambda = \lambda_\tau$ (resp. $\rho = \rho_\tau$), where λ_τ (resp. ρ_τ) denotes the left (resp. right) GNS-representation of $\mathfrak{C}$ associated with τ.*

Standard results in the spectral theory of C^*-algebras of type I yield the following:

Corollary 3.4. *There exists a unique positive Borel measure μ_{Pl} on $\hat{\mathfrak{C}}$, the Plancherel measure of $\mathcal{H}$, such that we have the following decomposition of τ in irreducible characters of $\mathfrak{C}$:*

$$\tau = \int_{\pi \in \hat{\mathfrak{C}}} \chi_\pi d\mu_{\mathrm{Pl}}(\pi). \tag{3.3}$$

4. The Plancherel measure

We will now address the problem to describe the spectrum $\hat{\mathfrak{C}}$ of $\mathfrak{C}$ and the Plancherel measure μ_{Pl}. The spectrum of $\mathfrak{C}$ is a rather complicated topological space. But it turns out that μ_{Pl}-almost everywhere it can be described by a simpler structure, namely a compact orbifold. This orbifold is represented by (in the sense of [50]) a groupoid of unitary standard induction data $\mathcal{W}_{\Xi_u}$ which is canonically associated with the affine Hecke algebra $\mathcal{H}$ (see [56], [22]). Its space of objects Ξ_u consists of induction data of $\mathcal{H}$ and the arrows $\mathcal{W}_{\Xi_u}$ are twisting isomorphisms between induction data. We will exhibit an explicit (up to some positive real multiplicative constants which are independent of the base $\boldsymbol{q}$)) positive measure μ on the compact orbifold $|\Xi_u| := \mathcal{W}\backslash\Xi_u$ such that a suitable open dense subset of $(|\Xi_u|, \mu)$ with a complement of measure 0 describes $(\hat{\mathfrak{C}}, \mu_{\mathrm{Pl}})$ almost everywhere.

The method in [56] to find this almost explicit Plancherel formula is a calculation of residues, starting from a basic complex analytic representation of τ as an integral over a certain rational n-form with values in the linear dual of $\mathcal{H}$, over a coset $pT_u \subset T$ of the compact real form T_u with $p \in T_{\mathrm{rs}} := \mathrm{Hom}(X, \mathbb{R}_+)$ far in the negative chamber [55]. Although such residue computations are certainly not new (see e.g. [1], [2], [38], [49]), the treatment of the uniqueness of residue data is new and is based on a simple geometric lemma in distribution theory which goes back to joint work with Gert Heckman [27]. This improved treatment of the residues is surprisingly powerful. It is sufficient to compute the Plancherel measure of the center $\mathcal{Z}$ of $\mathcal{H}$ explicitly, and in particular the central projection of the support of the Plancherel measure follows exactly [56] (see also [30]). In combination with Lusztig's results on the structure of completions of affine Hecke algebras at central characters (see [39]) we reorganize in [56] the residues according to parabolic induction and we *derive* the Maass–Selberg relations, the unitarity of the normalized intertwining operators, and finally the explicit (up to positive real factors) product formula for the Plancherel density.

4.1. The discrete series representations. Let us first recall the definition of the discrete series and of tempered representations:

Definition 4.1. An irreducible representation (V, π) of $\mathcal{H}$ is called a *discrete series representations* if it is equivalent to a subrepresentation of $(L^2(\mathcal{H}), \lambda)$. Equivalently, (V, π) is a discrete series representation if its character χ_π extends continuously to $L^2(\mathcal{H})$.

Remark 4.2. As an immediate consequence of this definition, a discrete series representation (V, π) can be equipped with an Hermitian inner product $\langle \cdot, \cdot \rangle$ with respect to which $\pi(h^*) = \pi(h)^*$ for all $h \in \mathcal{H}$. Such a Hilbert space representation of $\mathcal{H}$ is called unitary.

We will describe an *algebraic criterion* for a central character $W_0 t \in W_0 \backslash T$ to be the central character of a discrete series representation. For this we need to introduce the Macdonald c-function (see [45], [55]). This c-function is introduced as an element of the field of fractions of $\mathcal{A}$. The ring $\mathcal{A}$ can be interpreted as the ring of regular functions on T via $\theta_x \to x$, and thus the c-function can be interpreted as a rational function on T. Explicitly, we put

$$c := \prod_{\alpha \in R_{1,+}} c_\alpha, \tag{4.1}$$

where c_α is defined for $\alpha \in R_1$ by

$$c_\alpha := \frac{(1 + q_{\alpha^\vee}^{-1/2} \theta_{-\alpha/2})(1 - q_{\alpha^\vee}^{-1/2} q_{2\alpha^\vee}^{-1} \theta_{-\alpha/2})}{1 - \theta_{-\alpha}}. \tag{4.2}$$

The square roots here are positive square roots; observe that this formula makes sense: if $\alpha/2 \notin X$ then we have $q_{2\alpha^\vee} = 1$ (since $2\alpha^\vee \notin R_{\mathrm{nr}}$) and thus the numerator reduces to $(1 - q_{\alpha^\vee} \theta_{-\alpha})$.

We remark that there is no problem in defining the pole order of the rational function

$$\nu(t) := (c(t)c(t^{-1}))^{-1} \tag{4.3}$$

at a point $t_0 \in T$, since $\nu(t)$ is equal to a product of the rational functions of the form $(c_\alpha(t)c_\alpha(t^{-1}))^{-1}$ (with $\alpha \in R_1$). This function is the pull back via $\alpha/2$ (or α) of a rational function on $\mathbb{C}^\times$ and so it has a well defined pole order at t_0. The pole order of $\nu(t)$ at t_0 is defined as the sum of these pole orders.

The following theorem is of crucial importance.

Theorem 4.3 ([56, Corollary A.12]). *For any point $t_0 \in T$, the pole order of $\nu(t)$ at t_0 is at most equal to the rank* $\mathrm{rk}(R_0)$ *of R_0.*

Definition 4.4. We call $t_0 \in T$ a *residual point* if the pole order of $\nu(t)$ at t_0 is equal to the rank $\mathrm{rk}(X)$ of X.

Theorem 4.3 was proved in [56] by reducing it to a case by case inspection using the classification of residual points for graded Hecke algebras in [27], Section 4 [1]. The following result follows easily from Theorem 4.3.

Corollary 4.5. *For any root datum $\mathcal{R}$ and positive real label function q the set of residual points in T is a finite union of W_0-orbits. This set is nonempty only if* $\mathrm{rk}(R_0) = \mathrm{rk}(X)$.

Example 4.6 (The split adjoint case). *By the split adjoint case we mean that $f_s = 1$ for all $s \in S^{\mathrm{aff}}$ and $X = P$. The work of Kazhdan and Lusztig* [34] *implies* (*see* [56]*, Appendix B for the translation*) *that the residual points are the points of T of the following form. Let G be the Langlands dual group, i.e. the complex semisimple group with root datum $\mathcal{R}$* (*G is simply connected*)*. Then T is a maximal torus for G. Let $s \in T_u$ be such that $G_s \subset G$ is semisimple. Let $\mathcal{O}$ be a distinguished unipotent orbit of G_s and choose a homomorphism $\phi\colon \mathrm{SL}_2(\mathbb{C}) \to G_s$ with the property that*

$$\phi\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathcal{O} \quad \text{and} \quad c := \phi\begin{pmatrix} \boldsymbol{q}^{1/2} & 0 \\ 0 & \boldsymbol{q}^{-1/2} \end{pmatrix} \in T.$$

Then $r = sc \in T$ is a residual point, and all residual points in this case are of this form.

In general there exists an effective algorithm to classify the set of residual points for any root datum $\mathcal{R}$ with indeterminate positive real label function q (see [56, Theorem A.7]). The residual points come in finitely many generic families of residual points:

[1] The notion of a residual point in [27] seems more restrictive at first sight since it involves the existence of a full flag of intermediate "residual subspaces". In [56], Lemma A.11 we show however that this technical condition is always fulfilled. The existence of such full flags is the main tool for the classification of residual points in [27].

Proposition 4.7. *Let $\mathcal{R}$ be a semisimple root datum. There exists a nonempty finite W_0-invariant set* Res *of generic residual points $r\colon \mathcal{Q} \to T$ of $\mathcal{R}$. If $r \in$ Res then $r(q) = s.c(q)$ where $s \in T_u$ (T_u denoting the compact real form of T) is independent of q, and such that* $\mathrm{rank}(R_{0,s}) = \mathrm{rank}(R_0)$, *and where $c\colon \mathcal{Q} \to T_{\mathrm{rs}}$ (T_{rs} being the connected component of the split real form of T) is a group homomorphism such that for all $\alpha \in R_0$, $\alpha(c)^2$ is a monomial in the root labels $q_{\beta^\vee}$ ($\beta \in R_{\mathrm{nr}}$). For each generic residual point $r \in$ Res there exists an open set $\mathcal{Q}_r \subset \mathcal{Q}$ (depending only on the orbit $W_0 r$) which is the complement of finitely many (rational) hyperplanes in $\mathcal{Q}$, such that $r(q_0)$ is residual iff $q_0 \in \mathcal{Q}_r$ (see* [56, *Theorem A*.14]).

Remark 4.8. For each $\mathcal{R}$ one can explicitly determine the generic residual points r and the sets $\mathcal{Q}_r$. From the classification one can check that all residual points $r \in T$ of $\mathcal{R}$ have the important property that $\overline{r^{-1}} \in W_0 r$.

The following theorem expresses the central support of the discrete series representations of $\mathcal{H}$ in terms of residual points of T.

Theorem 4.9 ([56, Theorem 3.29]). *An orbit $W_0 r \in W_0\backslash T$ is the algebraic central character of a discrete series representation of $\mathcal{H}$ if and only if $r \in T$ is a residual point. In particular, the set $\Delta_{\mathcal{R}}$ of equivalence classes of discrete series representations of $\mathcal{H}$ is finite, and is nonempty only if* $\mathrm{rk}(R_0) = \mathrm{rk}(X)$.

As a consequence of the residue calculus one obtains an almost explicit *product formula* for the formal dimension (the Plancherel mass) of the discrete series as a function of the base $\boldsymbol{q}$.

Theorem 4.10 ([56, Corollary 3.32, Theorem 5.6]). *In this theorem we fix $f_s \in \mathbb{R}$ and we denote the corresponding half line in $\mathcal{Q}$ by $\mathcal{L} \subset \mathcal{Q}$ (see Remark* 2.3).

Notice that for each $r \in$ Res we have either $\mathcal{L} \subset \mathcal{Q}_r$ or $\mathcal{L} \cap \mathcal{Q}_r = \emptyset$. Let $r \in$ Res be such that $\mathcal{L} \subset \mathcal{Q}_r$. Via scaling isomorphisms [56, *Theorem* 5.6] *there exists a finite (trivial) fibration $\Delta_{W_0 r} \to \mathcal{L}$ whose fiber at $q \in \mathcal{L}$ is $\Delta_{W_0 r(q)}$, the finite set of discrete series representations of $\mathcal{H}(\mathcal{R}, q)$ with algebraic central character $W_0 r(q)$.*

Recall that we view $\boldsymbol{q} > 1$ as coordinate on $\mathcal{L}$. The expressions $\alpha(r) = \alpha(r(q))$ and $q_{\alpha^\vee}$ ($\alpha \in R_{\mathrm{nr}}$) with $q \in \mathcal{L}$ are thus viewed as functions of $\boldsymbol{q} > 1$. For each connected component $\delta \subset \Delta_{W_0 r}$ there exists a nonzero real constant $d_\delta \in \mathbb{R}^\times$ independent of $\boldsymbol{q}$ *such that for all $q \in \mathcal{L}$,*

$$\mu_{\mathrm{Pl}}(\{\delta(q)\}) = d_\delta \frac{q(w_0) \prod'_{\alpha \in R_1} (\alpha(r) - 1)}{\prod'_{\alpha \in R_1} (q_{\alpha^\vee}^{1/2} \alpha(r)^{1/2} + 1) \prod'_{\alpha \in R_1} (q_{\alpha^\vee}^{1/2} q_{2\alpha^\vee} \alpha(r)^{1/2} - 1)} \tag{4.4}$$

where $\prod'$ denotes the product in which we omit the factors which are equal to 0.

Remark 4.11. (a) Observe that only the constant d_δ depends on δ; the other factors only depend on the central character $W_0 r$ of δ.

(b) It was shown by Mark Reeder [57] that this leads to an effective way to determine L-packets of square integrable unipotent representations for exceptional

p-adic Chevalley groups (also see [28]) by considering the almost explicit product formulae for the Plancherel densities of the unipotent representations as a function of $\boldsymbol{q}$, the cardinality of the residue field [2]. Lusztig [41], [42] has given a more general parameterization of the unipotent L packets from a different point of view. In the cases considered the partitions of the set of square integrable unipotent representations coincide [57].

(c) This formula is explicit up to the nonzero real constant d_δ. We remark that if the restrictions $\chi_\delta|_{\mathcal{A}}$ of the characters χ_δ with $\delta \in \Delta_{W_0 r}$ are linearly independent then $d_\delta \in \mathbb{Q}$ for all $\delta \in \Delta_{W_0 r}$ ([56, Remark 3.35]). However we do not know this linear independence, and in fact for the more general class of tempered representations of $\mathcal{H}(\mathcal{R}, q)$ it is easy to find counterexamples for this. In any case, we conjecture that $d_\delta \in \mathbb{Q}$ (see [56], Conjecture 2.27).

(d) The constants d_δ were computed for the exceptional root systems (in the case $q(s) = \boldsymbol{q}$ for all $s \in S^{\mathrm{aff}}$, and $X = P$) in terms of the Kazhdan–Lusztig parameters of δ by Mark Reeder [57].

(e) An irreducible representation of $\mathcal{H}$ is a discrete series if and only if its matrix coefficients have exponential decay with respect to the norm function $\mathcal{N}$ on W (see Definition 5.1); this follows from the Casselman conditions (see [56, Lemma 2.22, Theorem 6.1(ii)]). In particular these matrix coefficients belong to the Schwartz algebra $\mathcal{S} \subset \mathfrak{C}$ (see Definition 5.1), and this shows that the discrete series are isolated points in $\hat{\mathfrak{C}}$.

4.2. Standard parabolic structures. We will now concentrate on the higher dimensional spectral series. Let $\mathcal{P}$ denote the power set of F_0. Let $R_P \subset R_0$ be the root subsystem $\mathbb{R}P \cap R_0$. Notice that P is a basis of simple roots for R_P. With $P \in \mathcal{P}$ we associate the sub root datum $\mathcal{R}^P := (X, R_P, Y, R_P^\vee, P)$. The associated nonreduced root system $R_{\mathrm{nr}}(\mathcal{R}^P)$ is equal to $R_{\mathrm{nr}}(\mathcal{R}^P) = \mathbb{R}P \cap R_{\mathrm{nr}}$ hence we can restrict the root labels on R_{nr} to $R_{P,\mathrm{nr}}$. This defines a label function q^P for the affine Weyl group of the root datum $\mathcal{R}^P$. We now define the subalgebra

$$\mathcal{H}^P := \mathcal{H}(\mathcal{R}^P, q^P) \hookrightarrow \mathcal{H}. \tag{4.5}$$

This affine Hecke algebra will typically not be semisimple. Its semisimple quotient is called $\mathcal{H}_P$, the quotient of $\mathcal{H}^P$ by the two sided ideal generated by $\theta_x - 1$ where $x \in Z^P := \{x \in X \mid x(\alpha^\vee) = 0 \forall \alpha \in P\}$. Notice that the elements θ_x with $x \in Z^P$ are central in $\mathcal{H}^P$. Let us denote by X_P the quotient $X_P = X/Z^P$, and by $Y_P = Y \cap \mathbb{R}P^\vee \subset Y$ its dual lattice. This gives us a semisimple root datum $\mathcal{R}_P := (X_P, R_P, Y_P, R_P^\vee, P)$. Again we see that $R_{\mathrm{nr}}(\mathcal{R}_P) = \mathbb{R}P \cap R_{\mathrm{nr}}$, so from the restriction of the root labels on R_{nr} to $R_{\mathrm{nr}}(\mathcal{R}^P)$ we can define a label function q_P on

[2] At the time when [57] was written, (4.4) was only conjectural for general discrete series representations (see [28]). Instead Reeder used a general Euler–Poincaré type formula for the formal dimension of a discrete series representation of a semisimple p-adic group (see [59]), which requires case-by-case considerations and presents serious computational difficulties. With (4.4) at hand some of these aspects of [57] can be simplified.

the affine Weyl group of $\mathcal{R}_P$. It is now easy to check that the semisimple quotient $\mathcal{H}_P$ of $\mathcal{H}^P$ is equal to

$$\mathcal{H}^P \twoheadrightarrow \mathcal{H}_P = \mathcal{H}(\mathcal{R}_P, q_P). \tag{4.6}$$

4.3. Twisting. Let $T_P = \mathrm{Hom}(X_P, \mathbb{C}^\times) \subset T$ be the character torus of X_P. Let T^P be the connected component of e of the subgroup $\{t \in T \mid \alpha(t) = 1\} \subset T$. We see that $T = T_P T^P$, that $K_P := T_P \cap T^P$ is a finite abelian group, and that T^P is pointwise fixed by the action of W_P on T.

Using the cross relations (2.5) we check that $t \in T^P$ gives rise to an automorphism

$$\phi_t : \mathcal{H}^P \to \mathcal{H}^P, \quad \phi_t(\theta_x N_w) = t(x)\theta_x N_w \tag{4.7}$$

of $\mathcal{H}^P$ (where $w \in W_P$).

Definition 4.12. Let Δ_P denote the set of equivalence classes of discrete series representations of $\mathcal{H}_P$. For $\delta \in \Delta_P$ and $t \in T^P$ we write $\delta_t = \tilde{\delta} \circ \phi_t$ for the twist by ϕ_t of the lift $\tilde{\delta}$ of δ to $\mathcal{H}^P$. Observe that for $k \in K_P$, ϕ_k descends to an automorphism ψ_k of $\mathcal{H}_P$.

Let $\mathfrak{W}_{P,P'} = \{w \in W_0 \mid w(P) = P'\}$. If $w \in \mathfrak{W}_{P,P'}$ then w induces an isomorphism of root data $w : \mathcal{R}^P \to \mathcal{R}^{P'}$ and $w : \mathcal{R}_P \to \mathcal{R}_{P'}$ which is compatible with the label functions. This defines corresponding isomorphisms

$$\phi_w : \mathcal{H}^P \to \mathcal{H}^{P'}; \quad \psi_w : \mathcal{H}_P \to \mathcal{H}_{P'} \tag{4.8}$$

of affine Hecke algebras.

Definition 4.13. Let $\delta \in \Delta_P$. We denote by $\delta^w = \delta \circ \psi_w^{-1} \in \Delta_{P'}$ the twist of δ by the isomorphism ψ_w. We define $(\delta_t)^w$ (with $t \in T^P$, $w \in \mathfrak{W}_{P,P'}$) similarly. Observe that $(\widetilde{\delta^w})_{wt} = (\delta_t)^w$. We denote by $\delta^k = \delta \circ \psi_k^{-1} \in \Delta_P$ the twist of δ by the automorphism ψ_k of $\mathcal{H}_P$. Observe that $(\widetilde{\delta^k})_{kt} = \delta_t$ if $k \in K_P$ and $t \in T^P$.

4.4. The groupoid of standard induction data. First of all, we denote by $\mathfrak{W}$ the (standard) Weyl groupoid, the groupoid whose set of objects is $\mathcal{P}$ and whose space of arrows from P to P' is equal to $\mathfrak{W}_{P,P'}$. This groupoid acts in an obvious way on the groupoid $\mathcal{K}$ whose set of objects is also $\mathcal{P}$ and whose space of arrows is described by $\mathcal{K}_{P,P'} = \emptyset$ if $P \neq P'$, and $\mathcal{K}_{P,P} = K_P$. We denote by $\mathcal{W}$ the semidirect product $\mathcal{W} = \mathcal{K} \rtimes \mathfrak{W}$, i.e. $\mathcal{W}$ is the finite groupoid whose set of objects is $\mathcal{P}$ and whose set of arrows from P to P' equals $\mathcal{W}_{P,P'} = K_{P'} \times \mathfrak{W}_{P,P'}$. The composition of arrows is given by $(k \times u)(l \times v) = ku(l) \times uv$.

The above twisting action on induction data coming from $\mathcal{H}_P$ naturally gives rise to an action of the groupoid $\mathcal{W}$ on the set Ξ of induction data of the various subquotient algebras $\mathcal{H}_P$ with $P \in \mathcal{P}$. Let Ξ be the set of triples (P, δ, t) with $P \in \mathcal{P}$, $\delta \in \Delta_P$ and $t \in T^P$. We see that Ξ is a finite union of complex algebraic tori of the form

$\Xi_{(P,\delta)} = \{(P, \delta, t) \mid t \in T^P\}$. In particular Ξ is fibered over $\mathcal{P}$ in a way compatible with the twisting action of $\mathcal{W}$.

The groupoid of standard induction data $\mathcal{W}_\Xi$ is the translation groupoid arising from the action of $\mathcal{W}$ on Ξ. Explicitly, if $\xi = (P, \delta, t)$, $\xi' = (P', \delta', t') \in \Xi_u$ then the set of arrows $\mathcal{W}_{\xi,\xi'}$ in $\mathcal{W}_\Xi$ between these induction data consists of the set of $g = k \times w \in K_{P'} \times \mathfrak{W}_{P,P'}$ such that $P' = w(P)$, $\delta' = \delta^g$, and $t' = gt$. One easily verifies that this forms an orbifold groupoid in the sense of [50]. We remark that the action of $\mathcal{K}$ on Ξ is free. Thus the quotient $\mathcal{W}_\Xi \to \mathcal{K}\backslash\mathcal{W}_\Xi = \mathfrak{W}_{\mathcal{K}\backslash\Xi}$ is a Morita equivalence and defines the same orbifold structure on $|\Xi| = \mathcal{W}\backslash\Xi$.

We denote by $\mathcal{W}_{\Xi_u}$ the full subgroupoid whose set of objects Ξ_u consists of the unitary standard induction data, i.e. the induction data of the form (P, δ, t) with $t \in T_u^P$, the compact real form of T^P. Hence Ξ_u is a finite disjoint union of compact tori, and $\mathcal{W}_{\Xi_u}$ is a compact orbifold groupoid.

4.5. The induction-intertwining functor. *We choose explicit representatives* (V_δ, δ) *for the equivalence classes* $\delta \in \Delta = \coprod_{P\in\mathcal{P}} \Delta_P$. Given $\xi = (P, \delta, t) \in \Xi$ we denote by $\pi(\xi)$ the induced representation $\mathrm{Ind}_{\mathcal{H}^P}^{\mathcal{H}}(\delta_t)$, realized on the finite dimensional vector space

$$i(V_\delta) = \mathcal{H} \otimes_{\mathcal{H}^P} V_\delta = \bigoplus_{w\in W^P} N_w \otimes V_\delta \tag{4.9}$$

where W^P denotes the set of shortest length representatives in W_0 for the left cosets of $W_P = W(R_P)$. It is not very difficult to show (see [6] or [56]) that for $\xi \in \Xi_u$, the induced representation $\pi(\xi)$ is unitary with respect to the Hermitian inner product on $i(V_\delta)$ defined by (with $x, y \in W^P$ and $u, v \in V_\delta$)

$$\langle N_x \otimes u, N_y \otimes v\rangle = \delta_{x,y}\langle u, v\rangle. \tag{4.10}$$

Theorem 4.14 ([56, Theorem 4.38]). *The assignment* $\Xi_u \ni \xi \to \pi(\xi)$ *extends to a functor* π *(the "induction intertwining" functor) from* $\mathcal{W}_{\Xi_u}$ *to* $\mathbb{P}\,\mathrm{Rep}(\mathcal{H})_{\mathrm{unit}}$*, the category of unitary modules of* $\mathcal{H}$ *in which the morphisms are unitary* $\mathcal{H}$*-intertwiners modulo scalars. This functor assigns a projectively unitary intertwining isomorphism* $\pi(g, \xi)\colon \pi(\xi) \to \pi(\xi')$ *to each* $g \in \mathcal{W}_{\xi,\xi'}$*. For all* $h \in \mathcal{H}$*, the map* $\xi \to \pi(\xi)(h)$ *extends to a regular function on* Ξ*, and for all* $g \in \mathcal{W}_{\xi,\xi'}$ *the map* $\xi \to \pi(g, \xi)$ *is rational but regular at* Ξ_u*. The functor* π *is independent of the choices of the realizations* (V_δ, δ) *of the* $\delta \in \Delta$ *up to natural isomorphisms.*

Remark 4.15. In fact the representations $\pi(\xi)$ with $\xi \in \Xi_u$ are known to be tempered, see below (see [56, Proposition 4.20]).

The projective representation π of $\mathcal{W}_{\Xi_u}$ canonically determines a 2-cohomology class $[\eta] \in H^2(\mathcal{W}_{\Xi_u}, S^1)$ of the groupoid $\mathcal{W}_{\Xi_u}$, namely the pull back via π of the 2-cohomology class of the standard central extension of the category of finite dimensional projective Hilbert spaces by S^1. Notice that $[\eta]$ is obviously a torsion class since the dimensions of the representations $\pi(\xi)$ are bounded by $|W_0|$.

In fact we can be a bit more precise here. Let $\mathcal{W}_\Delta$ be the finite groupoid which is defined like $\mathcal{W}_\Xi$ but with the finite set of objects Δ instead of Ξ, and its morphisms given by twisting. Then the assignment $(P,\delta)\to V_\delta$ can be upgraded to a projective representation of $\mathcal{W}_\Delta$ by *choosing*, for each arrow $g\in\mathcal{W}_\Delta^1$ with source δ, unitary intertwining isomorphisms $\delta_g^i\colon V_\delta\to V_{\delta^g}$ such that for all $h\in\mathcal{H}_{P'}$:

$$\delta_g^i\circ\delta(\psi_g^{-1}h)=\delta^g(h)\circ\delta_g^i. \tag{4.11}$$

One easily checks that such a choice defines a 2-cocycle η_Δ with values in S^1 by

$$\delta_g^i\circ(\delta')_{g'}^i=\eta_\Delta(g,g')(\delta')_{gg'}^i \tag{4.12}$$

where g, g' are composable arrows of $\mathcal{W}_\Delta$. Its class $[\eta_\Delta]\in H^2(\mathcal{W}_\Delta,S^1)$ is independent of the chosen representatives and intertwining morphisms. Then the details of the construction of the projective representation π actually show that $[\eta]\in H^2(\mathcal{W}_{\Xi_u},S^1)$ is the pull back of $[\eta_\Delta]\in H^2(\mathcal{W}_\Delta,S^1)$ via the natural homomorphism of groupoids

$$\mathcal{W}_{\Xi_u}\to\mathcal{W}_\Delta,\quad (P,\delta,t)\mapsto(P,\delta). \tag{4.13}$$

Let D denote the order of $[\eta_\Delta]$. Then we can choose the 2-cocycle η_Δ so that it has its values in μ_D, the group of complex D-th roots of unity. Then the above amounts to

Proposition 4.16. *Let $\tilde{\mathcal{W}}_{\Xi_u}$ denote the central extension of $\mathcal{W}_{\Xi_u}$ by μ_D determined by $[\eta]$. The lifting $\tilde\pi$ to $\tilde{\mathcal{W}}_{\Xi_u}$ of the projective representation π of $\mathcal{W}_{\Xi_u}$ splits.*

4.6. The Plancherel decomposition of τ. We finally have everything in place to formulate the Plancherel theorem for the spectral decomposition of $L^2(\mathcal{H})$ as a (type I) representation of $\mathfrak{C}$. In order to describe the Plancherel density, we introduce relative c-functions for the spectral series of the form $\pi(\xi)$ with $\xi\in\Xi_{(P,\delta),u}:=\{(P,\delta,t)\mid \delta\in\Delta_P, t\in T_u^P\}$.

Definition 4.17. We adopt the notation (P,α) to denote the restriction of $\alpha\in R_0\backslash R_P$ to $T^P\subset T$. Let X^P denote the character lattice of T^P. We write $R^P\subset X^P\backslash\{0\}$ for the set of restrictions (P,α) of roots $\alpha\in R_0\backslash R_P$ which are in addition primitive in the sense that if $\beta\in R_0\backslash R_P$ and $(P,\alpha)\in R^P$ such that (P,α) and (P,β) are proportional, then $(P,\beta)=c(P,\alpha)$ with $c\in\mathbb{Z}$. We write R_+^P for the primitive restrictions corresponding to the positive roots $\alpha\in R_{0,+}\backslash R_{P,+}$. An element (P,α) is called *simple* if (P,α) is indecomposable in $\mathbb{Z}_+R_+^P$. This is equivalent to saying that (P,α) is the restriction of an element of $F_0\backslash P$. To each $(P,\alpha)\in R^P$ we denote by $(P,H_\alpha)\subset T^P$ the connected component of e of $\mathrm{Ker}(P,\alpha)\subset T^P$. It is a codimension 1 subtorus of T^P. The real hyperplanes $(P,H_\alpha)\cap T_{\mathrm{rs}}^P$ are called the *walls* in T_{rs}^P. The positive chamber $T_{\mathrm{rs}}^{P,+}$ is the connected component of the complement of the walls on which $(P,\alpha)>1$ for all $(P,\alpha)\in R_+^P$.

Let $(P, \alpha) \in R^P$. Then the rational function on T defined by

$$c_{(P,\alpha)}(t) := \prod_{\beta \in R_{0,+} \setminus R_P : (P,\beta) \in \mathbb{Z}_+(P,\alpha)} c_\beta(t) \tag{4.14}$$

is clearly $W_P = W(R_P)$-invariant. Hence it can be viewed as a rational function on $W_P \backslash T$. We compose this rational function with the algebraic central character map $z_P \colon \mathrm{Irr}(\mathcal{H}^P) \to W_P \backslash T$ and write $c_{(P,\alpha)}(\sigma)$ for any $\sigma \in \mathrm{Irr}(\mathcal{H}^P)$. We define a rational function $c_{(P,\alpha)}$ on $\Xi_{(P,\delta)}$ by putting $c_{(P,\alpha)}(\xi) := c_{(P,\alpha)}(\delta_t)$. Observe that its poles and zeroes are orbits of (P, H_α) acting on $\Xi_{(P,\delta)}$.

Definition 4.18. (i) We define the c-function on $\xi \in \Xi_{(P,\delta)}$ by means of the formula $c(\xi) = \prod_{(P,\alpha) \in R_+^P} c_{(P,\alpha)}(\xi)$.

(ii) We put $\mu_{(P,\alpha)}(\xi) = (c_{(P,\alpha)}(\xi) c_{(P,-\alpha)}(\xi))^{-1} = |c_{(P,\alpha)}(\xi)|^{-2}$ (see Remark 4.8).

(iii) If $\xi \in \Xi_{(P,\delta)}$ we put $\mu(\xi) = \prod_{(P,\alpha) \in R_+^P} \mu_{(P,\alpha)}(\xi) = (c(\xi) c(w^P \xi))^{-1}$ (where w^P denotes the longest element in W^P).

The following main theorem of [56] makes the abstract Plancherel formula (3.3) almost (up to the constants $d_\delta \in \mathbb{R}_+$ for $\delta \in \Delta$) explicit.

Theorem 4.19 ([56, Theorem 4.39, Proposition 4.42, Theorem 4.43]). (i) *For all $(P, \alpha) \in R_+^P$ the functions $\mu_{(P,\alpha)}$ are smooth and nonnegative on $\Xi_{(P,\delta),u}$. Moreover μ is $\mathcal{W}$-invariant on Ξ.*

(ii) *The function $\Delta \ni (P, \delta) \to \mu_{(\mathcal{R}_P, q_P), \mathrm{Pl}}(\{\delta\})$ (given by the product formula (4.4) applied to $\mathcal{H}(\mathcal{R}_P, q_P)$) is $\mathcal{W}$-invariant on Δ.*

(iii) *Outside a $\mathcal{W}$-invariant subset Ξ_u^{reg} whose complement in Ξ_u consists of finitely many components of codimension at least 1, the representations $\pi(\xi)$ are irreducible. This defines a homeomorphism $[\xi] \to [\pi(\xi)]$ from $|\Xi_u^{\mathrm{reg}}| := \mathcal{W} \backslash \Xi_u^{\mathrm{reg}}$ onto a subset of $\hat{\mathfrak{C}}$ whose complement has measure zero.*

(iv) *For ξ in the compact torus $\Xi_{(P,\delta),u}$, let $d\xi$ denote the normalized Haar measure on $\Xi_{(P,\delta)}$. The Plancherel measure μ_{Pl} is the push forward to the quotient $|\Xi_u| = \mathcal{W} \backslash \Xi_u$ of the absolutely continuous, $\mathcal{W}$-invariant measure on Ξ_u given by (with $\xi = (P, \delta, t) \in \Xi_{(P,\delta),u}$)*

$$d\mu_{\mathrm{Pl}}(\pi(\xi)) = q(w^P)^{-1} |\mathfrak{W}_P|^{-1} \mu_{(\mathcal{R}_P, q_P), \mathrm{Pl}}(\{\delta\}) \mu(\xi) d\xi \tag{4.15}$$

where $\mathfrak{W}_P$ denotes the set $\{w \in W_0 \mid w(P) \subset F_0\}$.

5. The structure of the Schwartz algebra $\mathcal{S}$

In joint work with Patrick Delorme [22], [23] we studied the refinement of the Fourier isomorphism where the L^2-functions are replaced by smooth functions. Following Harish-Chandra and Langlands this step miraculously leads to deep insights in the structure of $\hat{\mathfrak{C}}$ and of $\mathrm{Irr}(\mathcal{H})$. It also brings into play methods from noncommutative geometry.

5.1. The Schwartz algebra $\mathcal{S}$. We define a sub-multiplicative function $\mathcal{N}: W \to \mathbb{R}_+$ as follows. Let $Z = X^+ \cap X^- \subset X$ be the lattice of central translations in W. Given $v \in \mathfrak{a}^*$ we denote by $\overline{v} \in \mathbb{R} \otimes Z$ the projection of v onto $\mathbb{R} \otimes Z$ along $\mathbb{R} \otimes Q(R_0)$. We choose a Euclidean norm $\| \cdot \|$ on $\mathbb{R} \otimes Z$ and put for any $w \in W$:

$$\mathcal{N}(w) := l(w) + \|\overline{w(0)}\|. \tag{5.1}$$

Now we come to the definition of the Schwartz space completion $\mathcal{S}$ of $\mathcal{H}$.

Definition 5.1. The Schwartz space completion $\mathcal{S} \subset L^2(\mathcal{H})$ of $\mathcal{H}$ is the nuclear Fréchet space consisting of the elements $\sum_{w \in W} c_w N_w \in L^2(\mathcal{H})$ such that $w \to |c_w|$ is of rapid decay with respect to the function $\mathcal{N}$.

As an application of the explicit knowledge on $\hat{\mathfrak{C}}$ we obtained in the previous section one can actually prove that

Theorem 5.2 ([56, Theorem 6.5]). *The multiplication of $\mathcal{H}$ extends continuously to $\mathcal{S}$, giving $\mathcal{S}$ the structure of a nuclear Fréchet algebra. Moreover, matrix coefficients of discrete series representations are elements of $\mathcal{S}$.*

5.2. Tempered representations

Definition 5.3. A finite dimensional representation (V, π) of $\mathcal{H}$ is called tempered if it extends continuously to $\mathcal{S}$. Equivalently, (V, π) is tempered if χ_π extends continuously to $\mathcal{S}$.

The next theorem explains the relation of tempered representations with L^2-theory:

Theorem 5.4 ([22, Theorem 3.19]). *If $\xi \in \Xi_u$ then $\pi(\xi)$ is unitary and tempered. Conversely, let (V, π) be an irreducible tempered $\mathcal{H}$ representation. Then there exists a $\xi \in \Xi_u$ such that (V, π) is a direct summand of $\pi(\xi)$.*

Corollary 5.5. *Irreducible tempered representations of $\mathcal{H}$ are unitarizable. The set $\hat{\mathcal{S}}$ of irreducible tempered representations is equal to the support $\hat{\mathfrak{C}}$ of the Plancherel measure.*

5.3. The Fourier isomorphism. Let us denote by $\mathcal{V}_{\Xi_u}$ the trivial fibre bundle over Ξ_u whose fibre at $\xi = (P, \delta, t) \in \Xi_u$ is $V_\xi := i(V_\delta)$ (the representation space of $\pi(\xi)$), thus

$$\mathcal{V}_{\Xi_u} := \coprod_{(P,\delta)} \Xi_{(P,\delta),u} \times i(V_\delta). \tag{5.2}$$

We denote by $\mathrm{End}(\mathcal{V}_{\Xi_u})$ the corresponding bundle of endomorphism algebras. Let $\mathrm{Pol}(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}))$ denote the algebra of polynomial sections in $\mathrm{End}(\mathcal{V}_{\Xi_u})$, i.e. sections such that the matrix coefficients are Laurent polynomials on each component $\Xi_{(P,\delta),u}$ of Ξ_u.

There is a natural action of $\mathcal{W}$ on $\mathrm{End}(\mathcal{V}_{\Xi_u})$ by algebra homomorphisms as follows. Suppose that $\xi \in \Xi_u$ and that $A \in \mathrm{End}(V_\xi)$. Given $g \in \mathcal{W}_{\xi,\xi'}$ we define $g(A) := \pi(g,\xi)\circ A\circ\pi(g,\xi)^{-1} \in \mathrm{End}(V_{\xi'})$. A section of f of $\mathrm{End}(\mathcal{V}_\Xi)$ is called $\mathcal{W}$-equivariant if we have $f(g(\xi)) = g(f(\xi))$ for all $\xi \in \Xi_u$ and all $g \in \mathcal{W}_{\Xi_u}$ with source ξ.

Observe that for all $h \in \mathcal{H}$, the polynomial section $\xi \to \pi(\xi)(h)$ on Ξ_u is $\mathcal{W}$-equivariant. Let us denote the algebra of $\mathcal{W}$-equivariant polynomial sections by $\mathrm{Pol}(\Xi_u, \mathrm{End}(\mathcal{V}_\Xi))^{\mathcal{W}}$. The Fourier transform on $\mathcal{H}$ is the canonical algebra homomorphism

$$\begin{aligned} \mathcal{F}_{\mathcal{H}} &: \mathcal{H} \to \mathrm{Pol}(\Xi_u, \mathrm{End}(\mathcal{V}_\Xi))^{\mathcal{W}}, \\ &\quad h \mapsto \{\xi \to \pi(\xi)(h)\}. \end{aligned} \tag{5.3}$$

The image of $\mathcal{F}_{\mathcal{H}}$ is difficult to describe. But it turns out that the situation becomes very intelligible upon extension to $L^2(\mathcal{H})$ or to $\mathcal{S}$.

Let $L^2(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}), d\mu_{\mathrm{Pl}})$ denote the Hilbert space of L^2-sections of $\mathrm{End}(\mathcal{V}_{\Xi_u})$ with respect to the inner product

$$\langle \sigma, \tau \rangle = \int_{\Xi_u} \mathrm{trace}(\sigma(\xi)^*\tau(\xi)) d\mu_{\mathrm{Pl}}(\pi(\xi)). \tag{5.4}$$

Let the Fréchet algebra of smooth sections of $\mathrm{End}(\mathcal{V}_{\Xi_u})$ on Ξ_u be denoted by $C^\infty(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}))$.

Theorem 5.6 ([56, Theorem 4.43], [22, Theorem 5.3]). (i) *The Fourier transform $\mathcal{F}_{\mathcal{H}}$ extends to an isometric isomorphism*

$$\mathcal{F} : L^2(\mathcal{H}) \to L^2(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}), d\mu_{\mathrm{Pl}})^{\mathcal{W}}. \tag{5.5}$$

(ii) *The Fourier transform $\mathcal{F}$ restricts to an isomorphism $\mathcal{F}_{\mathcal{S}}$ of Fréchet algebras*

$$\mathcal{F}_{\mathcal{S}} : \mathcal{S} \to C^\infty(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}))^{\mathcal{W}}. \tag{5.6}$$

5.4. The center of $\mathcal{Z}_{\mathcal{S}}$ of $\mathcal{S}$. As a consequence of Theorem 5.6 we see:

Corollary 5.7. *The center $\mathcal{Z}_{\mathcal{S}}$ of $\mathcal{S}$ is, via the Fourier Transform $\mathcal{F}_{\mathcal{S}}$, isomorphic to the algebra $C^\infty(\Xi_u)^{\mathcal{W}}$ of $\mathcal{W}$-invariant C^∞-functions on Ξ_u. In particular the algebra $\mathcal{S}$ is a $\mathcal{Z}_{\mathcal{S}}$-algebra of finite type.*

This gives, for irreducible tempered representations, a finer notion of central character which we call the *tempered central character*.

Definition 5.8. We denote by $\hat{\mathcal{S}}$ the set of irreducible tempered representations, equipped with its Jacobson topology.

It is easy to see that in our situation $\hat{\mathcal{S}}$ is naturally in bijection with the set $\mathrm{Prim}(\mathcal{S})$ if primitive ideals of $\mathcal{S}$. In view of Corollary 5.5 we have a bijection $\hat{\mathcal{S}} \to \hat{\mathfrak{C}}$, and it is not difficult to show that this bijection is in fact a homeomorphism.

Proposition 5.9. *We have a surjective, finite, continuous map*

$$z_{\mathcal{S}}\colon \hat{\mathcal{S}} \to |\Xi_u| = \mathcal{W}\backslash\Xi_u. \tag{5.7}$$

We call this map the tempered central character. *The irreducible summands of* $\pi(\xi)$ $(\xi \in \Xi_u)$ *all have central character* $[\xi] = \mathcal{W}_\xi \xi$.

5.5. Analytic R-groups. The structure of the fibres of the tempered central character map $z_{\mathcal{S}}$ is completely determined by the geometry of the orbifold $\mathcal{W}_{\Xi_u}$ in combination with the behaviour of the Plancherel density μ and the cohomology class η. This is revealed by studying the explicit inversion of the Fourier isomorphism $\mathcal{F}_{\mathcal{S}}$ by means of the wave packet operator $\mathcal{J}_{\mathcal{S}}$. Let $\mathcal{J}$ be the adjoint of $\mathcal{F}$ defined on the space $L^2(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}), d\mu_{\mathrm{Pl}})$ of non-invariant L^2-sections. Then $p_{\mathcal{W}} := \mathcal{F} \circ \mathcal{J}$ is equal to taking the $\mathcal{W}$-average, and [22, Theorem 5.3] shows that $p_{\mathcal{W}}$ maps the subspace of $L^2(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}), d\mu_{\mathrm{Pl}})$ consisting of sections of the form $c\sigma$ (with c the c-function and $\sigma \in C^\infty(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}))$ to the space $C^\infty(\Xi_u, \mathrm{End}(\mathcal{V}_{\Xi_u}))^{\mathcal{W}}$ of *smooth* invariant sections.

Definition 5.10. We call the connected components of the poles of $c_{(P,\alpha)}$ on $\Xi_{(P,\delta),u}$ the set of (P,α)-mirrors (since a root (P,α) is real, this set is empty or of codimension one in $\Xi_{(P,\delta),u}$). We define the set of mirrors on Ξ_u to be the union of all (P,α)-mirrors (with $P \in \mathcal{P}$ and $(P,\alpha) \in R^P$). This set is $\mathcal{W}$-invariant.

Using the above properties of $p_{\mathcal{W}}$ one shows that:

Proposition 5.11. (i) *For any* (P,α)*-mirror* $M \subset \Xi_{(P,\delta)}$ *there exists a mirror reflection* $\mathfrak{s}_M \in \mathcal{W}_{(P,\delta),(P,\delta)}$*, i.e. an involution in* $\mathcal{W}_{P,P}$ *fixing* δ *such that* $\mathfrak{s}_M$ *fixes* M *pointwise.*

(ii) *Let* $\xi \in \Xi_u$*, say* $\xi = (P,\delta,t)$*. The set* R_ξ *of roots* $(P,\alpha) \in R^P$ *such that* ξ *lies in a* (P,α)*-mirror is a reduced integral root system. We denote by* $R_{\xi,+}$ *the set of roots in* R_ξ *which are element of* $(P,\alpha) \in R^P_+$.

(iii) *For each mirror* $M \in \Xi_u$ *and for all* $\xi \in M$*, the intertwining operator* $\pi(\mathfrak{s}_M, \xi)$ *is scalar.*

Definition 5.12. Denote by $\mathcal{W}^m_\xi = W(R_\xi) \subset \mathcal{W}_{\xi,\xi}$ the normal reflection subgroup generated by the reflections in the mirrors through ξ. The analytic R-group $\mathfrak{R}_\xi$ at ξ is the group of $\mathfrak{r} \in \mathcal{W}_{\xi,\xi}$ such that $\mathfrak{r}(R_{\xi,+}) = R_{\xi,+}$. Then $\mathcal{W}_{\xi,\xi} = \mathcal{W}^m_\xi \rtimes \mathfrak{R}_\xi$.

The above implies that the restriction η_ξ of the 2-cocycle η to $\mathcal{W}_{\xi,\xi}$ is cohomologous to the pull-back of a 2-cocycle (also denoted η_ξ) of $\mathfrak{R}_\xi$. We have the following analog of results of Harish-Chandra ([25], [26]), Knapp–Stein ([36]) and Silberger [61] on analytic R-groups.

Theorem 5.13 ([22], [23]). *The restriction of the induction-intertwining functor* π *to* $\mathfrak{R}_\xi$ *gives rise to an isomorphism* $\mathrm{End}_{\mathcal{H}} V_\xi \simeq \mathbb{C}[\mathfrak{R}_\xi, \eta]$*, the* η_ξ*-twisted group ring of* $\mathfrak{R}_\xi$*. In particular, the functor* E_ξ *defined by* $\psi \to \mathrm{Hom}_{\mathbb{C}[\mathfrak{R}_\xi,\eta]}(\psi, V_\xi)$ *defines*

an equivalence between the category of $\mathbb{C}[\mathfrak{R}_\xi, \eta]$-modules and the category of finite dimensional unitary tempered $\mathcal{H}$-modules with tempered central character $\mathcal{W}\xi$.

Corollary 5.14. *Let $w \in \mathcal{W}_\xi$. There is a unique decomposition $w = xm$ with $m \in \mathcal{W}_\xi^m$ and $x(R_{\xi,+}) = R_{w\xi,+}$. Hence $x\mathfrak{R}_\xi x^{-1} = \mathfrak{R}_{w\xi}$. Consider the isomorphism of algebras $c_w : \mathbb{C}[\mathfrak{R}_\xi, \eta] \to \mathbb{C}[\mathfrak{R}_{w\xi}, \eta]$ defined by η-twisted conjugation with x, i.e. $c_w(r) = \eta(x, r)\eta(xr, x^{-1})\eta(x, x^{-1})^{-1} xrx^{-1}$. For any $\mathbb{C}[\mathfrak{R}_\xi, \eta]$-module ρ we have $E_\xi(\rho) = E_{w\xi}(\rho^w)$ where $\rho^w = \rho \circ c_w^{-1}$.*

Recall that the action of $\mathcal{K}$ on Ξ_u is free. We identify the R-groups $\mathfrak{R}_{k\xi}$ ($k \in \mathcal{K}$) via conjugation by k, and write $\mathfrak{R}_{\mathcal{K}\xi}$. We identify via the twisted conjugations c_k ($k \in \mathcal{K}$) the rings $\mathbb{C}[\mathfrak{R}_{k\xi}, \eta]$ ($k \in \mathcal{K}$) and write $\mathbb{C}[\mathfrak{R}_{\mathcal{K}\xi}, \eta_{\mathcal{K}}]$. We consider the sheaf $\mathrm{G}(\mathfrak{R}, \eta)$ on Ξ_u whose fiber at ξ is the complex representation space $\mathrm{G}(\mathfrak{R}_\xi, \eta)$ of $\mathbb{C}[\mathfrak{R}_\xi, \eta_\xi]$ (the argument that this defines a sheaf is the same as for the usual representation ring sheaf of the orbifold $\mathcal{W}_{\Xi_u}$, see [7]). By the above there exists a $\mathcal{W}$-sheaf $\underline{\mathrm{G}}(\mathfrak{R}, \eta)$ of complex vector spaces on Ξ_u (we identify $\mathfrak{R}_\xi$ with the quotient $\mathcal{W}_{\xi,\xi}/\mathcal{W}_\xi^m$, so that the action on characters by conjugation with $w \in \mathcal{W}_\xi$ is equal to c_w as in Proposition 5.14):

Definition 5.15. The $\mathcal{W}$-sheaf $\mathrm{G}(\mathfrak{R}, \eta)$ on Ξ_u descends to a sheaf of complex vector spaces on $|\Xi_u|$ denoted by $\underline{\mathrm{G}}(\mathfrak{R}, \eta)$, and to a sheaf $\mathrm{G}(\mathfrak{R}_{\mathcal{K}}, \eta_{\mathcal{K}})$ on $\mathcal{K}\backslash\Xi_u$.

6. Smooth families of tempered representations

We investigate the geometry of the orbifold structure of $|\Xi_u|$ in relation to the structure of the Grothendieck group of (finite dimensional) tempered representations of $\mathcal{H}$ in this subsection. The approach is comparable with [12] (but working with $\mathcal{S}$ instead of $\mathcal{H}$). Much of the material in this section and the next is joint work in progress with Maarten Solleveld.

As was explained by Arthur [3], it is a basic fact that the functor E_ξ is compatible with the geometric structure of $\mathcal{W}_{\Xi_u}$ in the following sense. The same arguments apply in the present situation. Let $\xi = (P, \delta, t)$ and let $P \subset Q$. Let $\mathfrak{R}_\xi^Q$ be the subgroup of elements of $\mathfrak{R}_\xi$ which pointwise fix the T_u^Q-orbit through ξ. Let π^Q denote the induction-intertwining functor of the Hecke algebra $\mathcal{H}^Q = \mathcal{H}(\mathcal{R}^Q, q^Q) \subset \mathcal{H}$, and let E_ξ^Q be the corresponding equivalence. Let Ξ_u^Q denote the space of unitary standard induction data for $\mathcal{H}^Q$. Then $\mathfrak{R}_\xi^Q$ is the R-group at $\xi \in \Xi_u^Q \subset \Xi_u$ for induction to $\mathcal{H}^Q$.

Proposition 6.1 ([3], [23]). *We have the following equality of functors:*

$$\mathrm{Ind}_{\mathcal{H}^Q}^{\mathcal{H}} \circ E_\xi^Q = E_\xi \circ \mathrm{Ind}_{\mathbb{C}[\mathfrak{R}_\xi^Q, \eta]}^{\mathbb{C}[\mathfrak{R}_\xi, \eta]}. \tag{6.1}$$

For any $\mathfrak{r} \in \mathfrak{R}_\xi$ the set of fixed points is of the form $T_u^Q \xi$ for some parabolic subsystem R_Q where Q is not necessarily standard, but *compatible* in the sense that there exists a $w \in \mathcal{W}_\xi$ such that $w(Q)$ is standard and $w(R_{\xi,+}) = R_{w\xi,+}$ (compare with [3, Section 2]). In combination with basic aspects of the structure theory of the (analytic or formal) completion of $\mathcal{H}$ at an algebraic central character (see [39] or [56]) we can draw several conclusions. Let $\mathrm{G}(\mathcal{S}_{\mathcal{W}\xi})$ denote the Grothendieck group of $\mathcal{S}$-modules with tempered central character $\mathcal{W}\xi$ where $\xi = (P, \delta, t)$, and let $\mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}\xi}) = \mathbb{Q} \otimes_{\mathbb{Z}} \mathrm{G}(\mathcal{S}_{\mathcal{W}\xi})$ (viewed as virtual tempered characters of $\mathcal{H}$).

Corollary 6.2. *Let $\chi \in \mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}\xi})$ where $\xi = (P, \delta, t) \in \Xi_{(P,\delta),u}$, and let $Q \in \mathcal{P}$ be such that $P \subset Q$. The following are equivalent:*

(i) *$\chi = E_\xi(\rho)$ with ρ induced from $\mathbb{C}[\mathfrak{R}_\xi^Q, \eta]$.*

(ii) *$\chi = E_\xi(\rho)$ with ρ supported on the $\mathfrak{R}_\xi$-conjugacy classes meeting $\mathfrak{R}_\xi^Q$.*

(iii) *χ is induced from a virtual tempered character of $\mathcal{H}^Q$.*

(iv) *$\chi = (\chi_t)|_{t=1}$ for a (weakly) smooth family $T_u^Q \ni t \to \chi_t \in \mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}(t\xi)})$.*

Proof. It is elementary (using the formula for induced characters in twisted group rings) that (i) $\Leftrightarrow$ (ii) and (i) $\Leftrightarrow$ (iii) follows from (6.1). It is trivial that (iii) $\Rightarrow$ (iv). The step (iv) $\Rightarrow$ (iii) needs explanation. By the (weak) smoothness of the family it enough to prove that χ_t is induced from $\mathcal{H}^Q$ for generic $t \in T_u^Q$. For generic $t \in T_u^Q$ the algebraic central character $z(\chi_t)$ is "R_Q-generic" in the sense of [56, Definition 4.12]. By [56, Corollary 4.15] all characters with an R^Q-generic central character are induced from $\mathcal{H}^Q$. □

Definition 6.3. We say that $\chi \in \mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}\xi})$ is *d-smooth and pure* if there exists an open neighborhood $U \subset \Xi_u$ of ξ, a d-dimensional smooth submanifold $S \subset U$ with $\xi \in S$, and a weakly smooth family $S \ni s \to \chi_s \in \mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}s})$ (weakly smooth means that $S \ni s \to \chi_s(h)$ is smooth for all $h \in \mathcal{H}$) such that the tempered central character of χ_s equals $\mathcal{W}s$ and $\chi = \chi_\xi$. We say that χ is *d-smooth* if χ is a $\mathbb{Q}$-linear combination of pure d-smooth characters.

We obtain a descending filtration $F^i_{\mathcal{W}\xi} = \mathbb{C} \otimes_{\mathbb{Q}} \{\chi \in \mathrm{G}^{\mathbb{Q}}(\mathcal{S}_{\mathcal{W}\xi}) \mid \chi \text{ is } i\text{-smooth}\}$ of $\mathrm{G}^{\mathbb{C}}(\mathcal{S}_{\mathcal{W}\xi})$. Observe that the filtration is stationary at least until the rank i_Z of the central lattice $Z \subset X$ of $\mathcal{R}$.

Definition 6.4. The vector space $\mathrm{Ell}^{\mathrm{temp}}_{\mathcal{W}\xi}$ of *elliptic tempered characters* of $\mathcal{S} = \mathcal{S}(\mathcal{R}, q)$ with tempered central character $\mathcal{W}\xi$ (with $\xi \in \Xi_u$) is defined as $F^0_{\mathcal{W}\xi}/F^1_{\mathcal{W}\xi}$. We call an irreducible tempered representation (with tempered central character $\mathcal{W}\xi$) elliptic if its character has a nonzero image in $\mathrm{Ell}^{\mathrm{temp}}_{\mathcal{W}\xi}$.

Proposition 6.5. *The complex vector space* $\mathrm{Ell}^{\mathrm{temp}}_{\mathcal{W}\xi}$ *is zero for all but finitely many orbits* $\mathcal{W}\xi \in |\Xi_u|$. *There exist orbits* $\mathcal{W}\xi \in |\Xi_u|$ *such that* $\mathrm{Ell}^{\mathrm{temp}}_{\mathcal{W}\xi} \neq 0$ *only if* $\mathcal{R}$ *is semisimple.*

One can show the following result:

Proposition 6.6. *Let* $F^d(\mathfrak{R}^Q_\xi, \eta)$ *denote the complex vector space of virtual* η*-twisted characters of* $\mathfrak{R}^Q_\xi$ *whose support consists of elements* $\mathfrak{r} \in \mathfrak{R}^Q_\xi$ *whose fixed point set in* $T_{Q,u}$ *has dimension at least* d. *In the case* $Q = F_0$ *this defines a filtration of* $\underline{\mathrm{G}}^{\mathbb{C}}_{\mathcal{W}\xi}(\mathfrak{R}, \eta)$ (*the complex span of irreducible* η*-twisted characters of* $\mathfrak{R}_\xi$). *Put* $I^Q_\xi := \mathrm{Ind}^{\mathbb{C}[\mathfrak{R}_\xi,\eta]}_{\mathbb{C}[\mathfrak{R}^Q_\xi,\eta]}$ *and define* $\mathrm{Ell}(\mathfrak{R}^Q_\xi, \eta) = F^0(\mathfrak{R}^Q_\xi, \eta)/F^1(\mathfrak{R}^Q_\xi, \eta)$. *We have an isomorphism of graded vector spaces*

$$\bigoplus_Q I^Q_\xi : \bigoplus_Q \mathrm{Ell}(\mathfrak{R}^Q_\xi, \eta) \xrightarrow{\sim} \mathrm{gr}(\underline{\mathrm{G}}^{\mathbb{C}}_{\mathcal{W}\xi}(\mathfrak{R}, \eta)) \tag{6.2}$$

where Q *runs over a complete set of representatives of the* $\mathcal{W}$*-association classes of compatible parabolic subgroups, and the degree of* $\mathrm{Ell}(\mathfrak{R}^Q_\xi, \eta)$ *is defined as the depth of* Q. *The functor* E_ξ *induces an isomorphism of graded vector spaces* $\mathrm{gr}(\underline{\mathrm{G}}^{\mathbb{C}}_{\mathcal{W}\xi}(\mathfrak{R}, \eta)) \to \mathrm{gr}(\underline{\mathrm{G}}^{\mathbb{C}}(\mathcal{S}_{\mathcal{W}\xi}))$. *If* $w \in \mathcal{W}_{\xi,\xi}$ *satisfies* $w(Q) = Q$ *then twisting by* w *acts trivially on* $\mathrm{Ell}(\mathfrak{R}^Q_\xi, \eta)$.

7. K-theory of the Schwartz algebra $\mathcal{S}$

In the last two sections we will discuss results and conjectures for the K-theory and noncommutative geometry of affine Hecke algebras and we indicate some of the evidence in support of these conjectures. These conjectures are not new, certainly not their analogues in the context of reductive algebraic groups. The point is however that these conjectures give a detailed guideline for representation theory of affine Hecke algebras with (unequal) continuous positive real parameters. In this sense the discussion below is a natural extension of the harmonic analysis questions that have been studied in this paper.

First of all recall the following result.

Proposition 7.1 (Corollary 5.9, [22]). *The dense* $*$*-subalgebra* $\mathcal{S} \subset \mathfrak{C}$ *is closed for holomorphic functional calculus. Hence the inclusion* $i : \mathcal{S} \to \mathfrak{C}$ *induces an isomorphism* $\mathrm{K}_*(\mathcal{S}) \xrightarrow{\sim} \mathrm{K}_*(\mathfrak{C})$.

7.1. The Chern character for $\mathcal{S}$. The result in this subsection is due to Maarten Solleveld. Proposition 7.1 shows that we can study the K-theory of the reduced C^*-algebra $\mathfrak{C}$ of $\mathcal{H}$ in terms of $\mathcal{S}$. By Theorem 5.6(ii) it follows easily that $\mathcal{S}$ is a Fréchet m-algebra, and for this type of topological algebras Cuntz [20] has shown the

existence of a unique functorial Chern character map $\mathrm{ch}\colon \mathrm{K}_* \to \mathrm{HP}_*$ with values in Connes' periodic cyclic homology HP_* and which is subject to certain natural compatibility properties. In [64] it is shown for a quite general class of Fréchet m-algebras that the Chern character becomes a natural isomorphism after tensoring with $\mathbb{C}$. Theorem 5.6(ii) shows that the Schwartz algebra $\mathcal{S}$ of $\mathcal{H}$ always falls in the class of algebras to which Solleveld's Theorem applies. Thus we have

Theorem 7.2 (Corollary 9, [64]). *The abelian group* $\mathrm{K}_*(\mathcal{S})$ *is finitely generated, and upon tensoring by* $\mathbb{C}$ *the Chern character* $\mathrm{Id}\otimes \mathrm{ch}\colon \mathrm{K}_*^{\mathbb{C}}(\mathcal{S}) \to \mathrm{HP}_*(\mathcal{S})$ *becomes an isomorphism, where we have used the notation* $\mathrm{K}_*^{\mathbb{C}}(\mathcal{S}) := \mathbb{C}\otimes_{\mathbb{Z}} \mathrm{K}_*(\mathcal{S})$.

7.2. Comparison between $\mathcal{S}$ and $\mathcal{H}$. The material this subsection is joint work in progress with Maarten Solleveld.

Conjecture 1 (Conjecture 8.9, [9]). The inclusion homomorphism $i\colon \mathcal{H} \to \mathcal{S}$ induces an isomorphism $\mathrm{HP}_*(\mathcal{H}) \to \mathrm{HP}_*(\mathcal{S})$.

There is quite solid evidence in support of this conjecture. Under certain additional assumptions we in fact have a proof of the statement (in fact, Solleveld has recently announced a general proof (private communication)). This proof is based on the philosophy explained in the important paper [35] and the Langlands classification for general affine Hecke algebras [23].

7.3. Independence of the parameters. Let us introduce the notation $\mathcal{S}(\boldsymbol{q}) := \mathcal{S}(\mathcal{R}, q)$ to stress the dependence on the base $\boldsymbol{q} > 1$ of the function q defined by $q(s) = \boldsymbol{q}^{f_s}$ while keeping the $f_s \in \mathbb{R}$ fixed. (In the terminology of Remark 2.3, we vary q in a half-line.) Let $\mathcal{S}_W = \mathcal{S}(1)$ denote the limiting case, the Schwartz algebra of functions $f\colon W \to \mathbb{C}$ of rapid decay. Observe that the Fréchet space $\mathcal{S}_W$ is isomorphic to $\mathcal{S}(\boldsymbol{q})$ as a Fréchet space via the naive linear isomorphism $f \to \sum_{w\in W} f(w)N_w$.

Solleveld has shown that there exists a family of isomorphisms (unique up to homotopy) $\psi_\varepsilon\colon \mathcal{S}(\boldsymbol{q}) \to \mathcal{S}(\boldsymbol{q}^\varepsilon)$ of pre C^*-algebras depending continuously on $\varepsilon \in (0, 1]$ such that $\psi_1 = \mathrm{id}_{\mathcal{S}}$. Now consider the family of linear isomorphisms $\phi_\varepsilon\colon \mathcal{S}_W \to \mathcal{S}(\boldsymbol{q})$ (with $\varepsilon \in (0, 1]$) consisting of the composition of the naive linear isomorphism $\mathcal{S}_W \to \mathcal{S}(\boldsymbol{q}^\varepsilon)$ with the inverse of ψ_ε. Solleveld has shown that the family ϕ_ε behaves as an asymptotic morphism $\mathcal{S}_W \to \mathcal{S}(\boldsymbol{q})$ as $\varepsilon \downarrow 0$ and defines a homomorphism $\mathrm{K}_*(\phi)\colon \mathrm{K}_*(\mathcal{S}_W) \to \mathrm{K}_*(\mathcal{S}(\boldsymbol{q}))$.

Conjecture 2. The map $\mathrm{K}_*^{\mathbb{Q}}(\phi)$ is an isomorphism.

This conjecture is the natural extension of [8, Conjecture 6.21] in the present context (after throwing away torsion), and in this sense it is entirely in the spirit of the Baum–Connes–Kasparov conjecture for p-adic reductive groups. In our situation (working with general continuous, unequal Hecke algebra parameters) it would be a very powerful principle for understanding the geometry of the irreducible spectrum

and the tempered irreducible spectrum of affine Hecke algebras, especially in combination with certain geometric refinements to be described below (see Conjectures 2b and 2c).

Let us first consider a weaker statement and the evidence in support of it:

Conjecture 2a (weaker version). $\dim(\mathrm{K}_*^{\mathbb{Q}}(\mathcal{S}(\boldsymbol{q}))) = \dim(\mathrm{K}_*^{\mathbb{Q}}(\mathcal{S}_W))$.

By Solleveld's Theorem 7.2 this statement is equivalent to $\dim(\mathrm{HP}_*(\mathcal{S}(\boldsymbol{q}))) = \dim(\mathrm{HP}_*(\mathcal{S}_W))$, and thus in cases where Conjecture 1 is known (certainly including the split case $f_s = 1, \forall s \in S^{\mathrm{aff}}$) it is equivalent to $\dim(\mathrm{HP}_*(\mathcal{H}(\boldsymbol{q}))) = \dim(\mathrm{HP}_*(\mathbb{C}[W]))$. In the split case this is a theorem of Baum and Nistor [10]. The proof in [10] is very interesting and is based on Lusztig's asymptotic morphism in combination with techniques from [35]. For unequal label Hecke algebras one may connect this with Lusztig's conjectures from [43], see [4].

7.4. Equivariant K-theory. The conjecture 2 can be expressed more geometrically in terms of equivariant K-theory using well known results (see [7], [10], [67]). First of all $\mathcal{S}_W = C^\infty(T_u) \rtimes W_0$, and this gives the identification $\mathrm{K}_*(\mathcal{S}_W) = \mathrm{K}_{W_0}^*(T_u)$. Recall that for any finite group G acting on a compact topological space X the equivariant Chern character defines an isomorphism $\mathrm{Id} \otimes \mathrm{ch}_{\mathrm{G}} : \mathbb{C} \otimes_{\mathbb{Z}} \mathrm{K}_G^*(X) \xrightarrow{\sim} \mathrm{H}^{[*]}(G\backslash\hat{X}, \mathbb{C})$ where $\hat{X} = \cup_{g\in G}(g, X^g)$ is the disjoint union of the fixed point spaces X^g of the elements of G, and where $\mathrm{H}^{[*]}$ denotes the $\mathbb{Z}/2\mathbb{Z}$-periodic Čech cohomology groups (see [7]). The topological space $G\backslash\hat{X}$ is called the extended quotient of X. It is the orbit space of the inertia orbifold $\Lambda(G_X) := G_{\hat{X}}$ of the translation orbifold $G_X = X \rtimes G$. In this geometric form the conjecture gives rise to a natural refinement of the conjecture for $\mathrm{K}_0(\mathcal{S})$ which is the main reason for this reformulation.

Definition 7.3. Given $\xi \in \mathcal{W}\backslash\Xi_u$ we have a quotient homomorphism $\pi_\xi : \mathcal{S} \to \mathcal{S}/m_\xi\mathcal{S}$ where m_ξ denotes the maximal ideal of $\mathcal{Z}_\mathcal{S}$ at ξ (the ring $\mathcal{S}/m_\xi\mathcal{S}$ is finite dimensional and semisimple by Theorem 5.6). For $\alpha \in \mathrm{K}_0(\mathcal{S})$ we define $\mathrm{Supp}(\alpha) = \{\mathcal{W}\xi \in \mathcal{W}\backslash\Xi_u \mid \mathrm{K}_0(\pi_\xi)(\alpha) \neq 0\}$ (a closed set in $\mathcal{W}\backslash\Xi_u$, as one checks easily).

Conjecture 2b (geometric refinement). There exists a natural isomorphism (take $\mathrm{K}_*^{\mathbb{C}}(\phi)$ composed with the inverse of the equivariant Chern character for the action of W_0 on T_u):

$$\kappa : \mathrm{H}^{[*]}(W_0\backslash\widehat{T_u}, \mathbb{C}) \xrightarrow{\sim} \mathrm{K}_*^{\mathbb{C}}(\mathcal{S}(\boldsymbol{q})). \tag{7.1}$$

The ascending filtration $F_i(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}(\boldsymbol{q}))) = \{\alpha \mid \dim(\mathrm{Supp}(\alpha)) \leq i\}$ of $\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}(\boldsymbol{q}))$ coincides via this isomorphism with the filtration of the left hand side whose i-th filtered piece consist of the sum of the even cohomologies of the components of the extended quotient $W_0\backslash\widehat{T_u}$ of dimension at most i.

This form of the conjecture guides us to a further reduction to the discrete part $F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}(\boldsymbol{q})))$ of the vector space $\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}(\boldsymbol{q}))$. This is best understood in the context of a conjecture concerning index theory.

8. Index functions

The material in this section was shaped in its present form in the course of various conversations with Mark Reeder and Joseph Bernstein. I am much indebted for their insightful comments. The ideas in this section have their origin in the theory of the Selberg trace formula. We want to construct "index functions" using the Euler–Poincaré principle in K-theory (see e.g. [37], [59], [21]).

We assume throughout in this section that the root datum $\mathcal{R}$ is semisimple unless stated otherwise. Every finite dimensional tempered module π gives rise to a $\mathbb{Z}$-valued "local index" function Ind_π on $\mathrm{K}_0(\mathcal{S})$. Namely $\pi\colon \mathcal{S} \to \mathrm{End}_{\mathbb{C}}(V_\pi)$ is a continuous algebra homomorphism, and given $\alpha = [p] \in \mathrm{K}_0(\mathcal{S})$ (with $p \in M_N(\mathcal{S})$ an idempotent) we define $\mathrm{Ind}_\pi(\alpha) = \mathrm{rank}(\pi(p)) \in \mathbb{Z}$. This naturally descends to bilinear pairing $[\,\cdot\,,\cdot\,]\colon \mathrm{K}_0(\mathcal{S}) \times \mathrm{G}(\mathcal{S}) \to \mathbb{Z}$ given by $[\alpha, [\pi]] := \mathrm{Ind}_{[\pi]}(\alpha)$. Moreover, by the structure theory of $\mathcal{S}$ (Theorem 5.6), it is clear that if the local index function $[\pi] \to \mathrm{Ind}_{[\pi]}(\alpha)$ vanishes identically on $\mathrm{G}(\mathcal{S})$ then $\alpha = 0$.

Now suppose that $\alpha \in F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}))$. By definition of the support of α and of our notion of smooth families this means that $\mathrm{Ind}_{[\pi]}(\alpha) = 0$ for all π which are 1-smooth (it must be constant along the family on the one hand, but on the other hand its support must be finite). Therefore $\pi \to \mathrm{Ind}_{[\pi]}(\alpha)$ factors through a function on $\mathrm{Ell}^{\mathrm{temp}}$, and $[\,\cdot\,,\cdot\,]$ factors through a bilinear pairing $[\,\cdot\,,\cdot\,]$ on $F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S})) \times \mathrm{Ell}^{\mathrm{temp}}$ which is non-degenerate on the left.

Next we consider the change of base ring homomorphism $\beta\colon \mathrm{K}_0(\mathcal{H}) \to \mathrm{K}_0(\mathcal{S})$ defined by $[P] \to [S \otimes_{\mathcal{H}} P]$ if P denotes a finitely generated projective $\mathcal{H}$-module. One would like to complement this base change homomorphism β with a base change homomorphism $\beta\colon \mathrm{K}(\mathrm{Mod}_{fg}(\mathcal{H})) \to \mathrm{K}(\mathrm{Mod}(\mathcal{S}))$ but there seems no obvious way to do this. First of all $\mathcal{S}$ is *not* flat over $\mathcal{H}$ (this problem already occurs in rank 1) and it is also not quite clear which category of $\mathcal{S}$-modules one should consider. The conjecture we are about to make precisely states that this can anyway be done if one restricts in some sense to the tempered modules of finite length. To this end we will first show that the projective dimension of $\mathcal{H}$-modules of finite length is bounded by the rank [3] (which is a joint result with Mark Reeder). Let (π, V) be an $\mathcal{H}$-module of finite length. Let $C \subset \mathfrak{a}^* = \mathbb{R} \otimes_{\mathbb{Z}} X$ be the fundamental alcove for the action of $W^a = W_0 \ltimes Q \lhd W$ (the normal subgroup generated by reflections in W). Let $\Omega \subset W$ be the finite abelian subgroup of elements of length 0, then $W = W^a \rtimes \Omega$. We denote by $\mathcal{H}^a \subset \mathcal{H}$ the unital subalgebra of $\mathcal{H}$ spanned by the elements $\{N_w\}_{w \in W^a}$, then $\mathcal{H} = \mathcal{H}^a \rtimes \Omega$ (a crossed product).

Given a nonempty facet $\emptyset \neq f \subset C$ of C we have the corresponding subset $S_f \subset S^{\mathrm{aff}}$ of affine simple reflections fixing f. Let $\mathcal{H}_f \subset \mathcal{H}^a$ be the finite type Hecke subalgebra $\mathcal{H}_f = \mathcal{H}(S_f, q_{S_f})$. For any subset $I \subset S^{\mathrm{aff}}$ we denote by $\mathbb{C}^I$ the

[3] One can deduce from this the fact that the category of $\mathcal{H}$-modules is of finite cohomological dimension, as was explained to me by Joseph Bernstein. One should compare this to Bernstein's result that the category of smooth representations of a reductive p-adic group has finite cohomological dimension, see [65, Prop. 37].

complex vector space which has as a basis the set I. Put

$$\tilde{C}_i(V) = \bigoplus_{f:\dim(f)=i} \mathcal{H} \otimes_{\mathcal{H}_f} (V|_{\mathcal{H}_f}) \otimes_{\mathbb{C}} \bigwedge^{n-i} \mathbb{C}^{S_f}. \tag{8.1}$$

Since $\mathcal{H}_f \subset \mathcal{H}$ is a finite dimensional semisimple subalgebra this is clearly a projective, finitely generated $\mathcal{H}$-module. We define $\mathcal{H}$-linear maps $\tilde{d}_i : \tilde{C}_i \to \tilde{C}_{i-1}$ by

$$\tilde{d}_i(h \otimes_{\mathcal{H}_f} v \otimes \lambda) := \bigoplus_{\substack{f' \subset f \\ \dim(f')=i-1}} h \otimes_{\mathcal{H}_{f'}} v \otimes (\lambda \wedge s_{f,f'}) \tag{8.2}$$

where $s_{f,f'} \in S^{\mathrm{aff}}$ is defined by $S_{f'} = S_f \cup \{s_{f,f'}\}$. Observe that there is a natural left Ω-action on $\tilde{C}_i(V)$ by means of $\mathcal{H}$-intertwining operators via

$$j_\omega(h \otimes_{\mathcal{H}_f} v \otimes \lambda) = h\omega^{-1} \otimes_{\mathcal{H}_{\omega(f)}} \pi(\omega)v \otimes \omega(\lambda). \tag{8.3}$$

This action commutes with the action of the operators $\tilde{d}_i$. Finally, we define $C_i(V) = (\tilde{C}_i(V))^{j(\Omega)}$ and we denote by d_i the restriction of $\tilde{d}_i$ to $C_i(V)$.

Proposition 8.1 (E. Opdam and M. Reeder, unpublished). *Let (V, π) be an $\mathcal{H}$-module of finite length. The graded $\mathcal{H}$-linear operator $d = \{d_i\}$ is a differential on $C_*(V)$, making $C_*(V)$ into a bounded complex of finitely generated projective $\mathcal{H}$-modules. The $\mathcal{H}$-linear map $d_0 : C_0(V) \to C_{-1}(V) \simeq V$ extends this to a finite projective resolution $0 \leftarrow V \xleftarrow{d_0} C_*(V)$ of V.*

The proof of this proposition reduces simply to the case $\mathcal{H} = \mathcal{H}^a$, and there the proof is a variation of Kato's proof [33] of a statement about a similar "restriction-induction" complex for finite type Hecke algebras.

Definition 8.2. By the previous result all $\mathcal{H}$-modules of finite length have finite projective dimension. Hence there is a well defined Euler–Poincaré homomorphism $\varepsilon : \mathrm{G}(\mathcal{H}) \to \mathrm{K}_0(\mathcal{H})$. It has an explicit realization $\varepsilon([V]) := \sum_i (-1)^i [C_i(V)]$.

Let $\rho : \mathrm{G}(\mathcal{S}) \to \mathrm{G}(\mathcal{H})$ be the homomorphism which corresponds to the forgetful functor (forgetting temperedness). We have now altogether constructed a homomorphism $\gamma = \beta \circ \varepsilon \circ \rho : \mathrm{G}(\mathcal{S}) \to \mathrm{K}_0(\mathcal{S})$. We extend this to an *anti-linear* map $\gamma : \mathrm{G}^{\mathbb{C}}(\mathcal{S}) \to \mathrm{K}_0^{\mathbb{C}}(\mathcal{S})$. It is not difficult to show that this map vanishes on modules induced from a proper parabolic subalgebra (e.g. by using the Koszul resolution for a regular sequence of parameters for the smooth family of induced representations). Thus we obtain an anti-linear map

$$\gamma : \mathrm{Ell}^{\mathrm{temp}} \to F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S})). \tag{8.4}$$

Using γ the previously defined bilinear pairing $[\,\cdot\,,\cdot\,]$ on $F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S})) \times \mathrm{Ell}^{\mathrm{temp}}$ gives rise to a sesquilinear form $\langle U, V\rangle_{\mathrm{ell}} := [\gamma(U), V]$ on $\mathrm{Ell}^{\mathrm{temp}}$. This is the

precise analog of the elliptic pairing of tempered characters as defined by Schneider and Stuhler [59]: suppose that U and V are finite dimensional tempered modules of $\mathcal{H}$, then

$$\langle U, V\rangle_{\mathrm{ell}} = [\gamma(U), V] = \sum_{i\geq 0}(-1)^i \dim \mathrm{Ext}^i_{\mathcal{H}}(U, V). \tag{8.5}$$

In this context we remark that the natural map $\mathrm{Ell}^{\mathrm{temp}} \to \mathrm{Ell}^{\mathrm{alg}}$ (the elliptic virtual representations of $\mathcal{H}$) is a linear isomorphism [23], by the Langlands parametrization [23] for affine Hecke algebras.

By the Euler–Poincaré principle applied to our standard resolution we can also express this explicitly (following [59], [58]) in terms of an "index function" for U. Let $\Omega_f \subset \Omega$ be the stabilizer of f in Ω, and let ε_f be the character of Ω_f on $\mathbb{C}^{S_f}$. We define the index function $f_U \in \mathcal{H}$ by

$$f_U = \sum_f (-1)^{\dim(f)} \sum_{\sigma \in \mathrm{Irr}(\mathcal{H}_f \rtimes \Omega_f)} \dim(\sigma)^{-1} [U|_{(\mathcal{H}_f \rtimes \Omega_f)} \otimes \varepsilon_f : \sigma] e_\sigma \in \mathcal{H} \tag{8.6}$$

where f runs over a complete set of representatives of the Ω-orbits of faces of C, and where $e_\sigma \in \mathcal{H}_f \rtimes \Omega_f$ denotes the central idempotent corresponding to σ in the finite dimensional complex semisimple algebra $\mathcal{H}_f \rtimes \Omega_f$. Then

$$\langle U, V\rangle_{\mathrm{ell}} = \chi_V(f_U) \tag{8.7}$$

By (8.5) it is clear that this pairing is Hermitian, and that (virtual) tempered representations U and V with distinct tempered central characters are orthogonal. Moreover, the pairing is integral with respect to the lattice generated by the elliptic (true) characters.

Conjecture 3. The pairing $\langle U, V\rangle_{\mathrm{ell}}$ on $\mathrm{Ell}^{\mathrm{temp}}_{W_\xi}$ corresponds, via the functor E_ξ of Theorem 5.13, with the elliptic paring on $\mathrm{Ell}(\mathfrak{R}_\xi, \eta)$ given by

$$\langle \phi, \chi\rangle_{\mathrm{ell}} = |\mathfrak{R}_\xi|^{-1} \sum_{\mathfrak{r}\in\mathfrak{R}_\xi} |\det(1-\mathfrak{r})| \overline{\phi(\mathfrak{r})} \chi(\mathfrak{r}) \tag{8.8}$$

(see [3] and [58]). In particular, this pairing is positive definite (since the support of the function $\det(1-\mathfrak{r})$ is exactly equal to the set of elliptic conjugacy classes of $\mathfrak{R}_\xi$).

This conjecture is the natural analog of results of Arthur in the theory of the local trace formula [3]. In the split case with $X = P$ it was shown by Mark Reeder [58] using the Kazhdan–Lusztig parameterization and a comparison between geometric and analytic R-groups. The formula (8.6) for the index functions f_U is due to Mark Reeder [57], based on work of Schneider and Stuhler [59]. Recently a related and very general result was obtained by R. Meyer [48] for Schwartz algebra's of reductive p-adic groups. In order to explain this, first observe that Theorem 5.6 implies that a discrete series representation (U, δ) of $\mathcal{S}$ is a projective $\mathcal{S}$-module. Therefore it

defines a class $[U]_{\mathcal{S}} \in K_0(\mathcal{S})$. On the other hand we have defined the class $\gamma([U]) \in K_0(\mathcal{S})$ above. In our context Meyer's result would mean that $[U]_{\mathcal{S}} = \gamma([U])$, and in particular that all higher extensions $\mathrm{Ext}^i_{\mathcal{H}}(U, V)$ $(i > 0)$ vanish if V is tempered (this is remarkable since $\mathcal{S}$ is not flat over $\mathcal{H}$). Meyer's Theorem actually implies the validity of this statement for all affine Hecke algebras which arise in connection with a reductive p-adic group G via the theory of types of equivalence classes of G-inertial cuspidal data (of course, we conjecture that it holds for general $\mathcal{H}$). This proves that in those cases Conjecture 3 holds for the discrete series representations of $\mathcal{H}$, thus providing strong evidence in support of this conjecture.

Corollary 8.3 (L^2-index for discrete series representations of $\mathcal{H}$). *We have the following explicit Euler–Poincaré formula for the formal dimension of a discrete series representation (U, δ) of $\mathcal{H}$, expressed in terms of its "K-types":*

$$\begin{aligned}\mu_{\mathrm{Pl}}(\{\delta(q)\}) &= \tau(f_U(q)) \\ &= \sum_f (-1)^{\dim(f)} \sum_{\sigma \in \mathrm{Irr}(\mathcal{H}_f \rtimes \Omega_f)} [U|_{(\mathcal{H}_f \rtimes \Omega_f)} \otimes \varepsilon_f : \sigma] d_\sigma(q) \end{aligned} \tag{8.9}$$

where f runs over a complete set of representatives of the Ω-orbits of faces of C, and where $d_\sigma(q)$ denotes the formal dimension of σ in the finite dimensional Hilbert algebra $\mathcal{H}_f \rtimes \Omega_f$ whose trace is the restriction of the trace τ of $\mathcal{H}$ (these are rational functions in the parameters $q_s^{1/2}$, $q_s^{-1/2}$ with rational coefficients).

Please compare this statement with the product formula (4.4) for the formal dimensions. Observe that the rationality of the constant d_δ in (4.4) is an immediate consequence. The Euler–Poincaré formula for f_U (as obtained by Schneider and Stuhler [59]) was used by Reeder [57] for the computation of all formal dimensions of the square integrable unipotent representations of Chevalley groups of exceptional type. The main computational work in [57] consists in the reduction of the Euler–Poincaré alternating sum to the product formula.

A second consequence of Conjecture 3 is:

Corollary 8.4. *The map $\gamma\colon \mathrm{Ell}^{\mathrm{temp}} \to F_0(\mathrm{K}_0^{\mathbb{C}}(\mathcal{S}))$ is an anti-linear isomorphism.*

Let us write $\mathrm{Ell}^{\mathrm{temp}}(q)$ in order to stress the dependence on q. Observe that $\mathrm{Ell}^{\mathrm{temp}}(q)$ is a semisimple $\mathcal{Z}$-module via the algebraic central character z map. The combination of Corollary 8.4 with Conjecture 2b implies that the dimension of the finite dimensional space (see Proposition 6.5) $\mathrm{Ell}^{\mathrm{temp}}(q)$ is independent of $q \in \mathcal{Q}$. We conjecture a stronger statement:

Conjecture 2c. The $\mathcal{Q}$-family of finite dimensional semisimple $\mathcal{Z}$-modules $q \to \mathrm{Ell}^{\mathrm{temp}}(q)$ is continuous (i.e. isomorphic to a direct sum of one-dimensional $\mathcal{Z}$-modules, each depending continuously on q).

Let us denote by $j\colon |\Xi_u| \to W_0\backslash T_u$ the map that sends $\mathcal{W}\xi$ (with $\xi = (P, \delta, t)$) to $W_0(|r|^{-1}rt)$ where $z_P(\delta) = W_P r$ is the central character of δ. Let $q = 1$ and let

$t \in T_u = \Xi_u(1)$. Then $\mathfrak{R}_t$ is the isotropy group $W_{0,t}$ of t in W_0. Using Proposition 6.6, Conjecture 2c implies an isomorphism of sheaves

$$j_*(\underline{\mathrm{G}}(\mathfrak{R}, \eta)) \simeq \underline{\mathrm{G}}(W_0) \tag{8.10}$$

on $W_0 \backslash T_u$, where the sheaf on the right hand side is the usual complex representation ring sheaf for the action of W_0 on T_u.

8.1. Discussion and examples. We discuss the implications of Conjecture 2c for the problem of understanding the tempered spectrum of non-simply laced affine Hecke algebras. The results in this section are joint with Maarten Solleveld. In these cases, Proposition 4.7, Theorem 4.9 and Conjecture 2c suggest to use deformations to generic points $q \in \mathcal{Q}$ in order to approach this problem.

Let us consider the three parameter case $\mathcal{R} = (\mathbb{Z}^n, B_n, \mathbb{Z}^n, C_n, F_0)$ with $F_0 = \{e_1 - e_2, \ldots, e_{n-1} - e_n, e_n\}$ (thus $Q(B_n) = \mathbb{Z}^n$; we refer to this case as "type C_n^{aff}") with parameters $q_0 = q(s_0)$, $q_1 = q(s_{e_i})$, $q_2 = q(s_{e_i - e_{i+1}})$. The case $n = 1$ will be considered as a special case (but with two parameters q_0, q_1); the discussion below applies to this degenerate case without modifications). The set of distinct W_0-orbits of generic residual points is easily seen to be parametrized (using [56, Appendix A]) by ordered pairs (μ, ν) of partitions of total weight n. The orbit of the unitary part of (μ, ν) only depends on $i := |\mu| \in \{0, 1, \ldots, n\}$. Let $W_0 s_i$ denote this orbit. The orbits $W_0 s_i \in W_0 \backslash T_u$ are mutually distinct, and the stabilizer group of s_i is isomorphic to the Weyl group of type $B_i \times B_{n-i}$.

We see that for each i the cardinality of the set of orbits of generic residual points whose corresponding orbit of unitary parts equals $W_0 s_i$ is precisely equal to the number of elliptic conjugacy classes of the stabilizer group $W(B_i) \times W(B_{n-i})$ of s_i, which is the R-group $\mathfrak{R}_{s_i}$ for $\mathcal{S}_W = \mathcal{S}(1)$. By Theorem 4.9 each orbit $W_0 r$ of residual points carries *at least one* discrete series representation of $\mathcal{H}$. We call $q \in \mathcal{Q}$ *generic* if $q \in \cap \mathcal{Q}_r$ (intersection over all $r \in \mathrm{Res}$) and if for all $r, r' \in \mathrm{Res}$: $W_0 r(q) = W_0 r'(q) \Rightarrow W_0 r = W_0 r'$. By Conjecture 2c and the above description of the set of orbits of generic residual points we conclude that for a generic parameter q each residual orbit $W_0 r(q)$ must carry *precisely one* discrete series representation. So for generic $q \in \mathcal{Q}$ the discrete series characters are separated by their algebraic central character. Moreover, the complex linear span of these discrete series characters is isomorphic to the space $\mathrm{Ell}^{\mathrm{temp}}(q)$.

By the continuity aspect of Conjecture 2c we are now in principle able to understand the spaces $\mathrm{Ell}^{\mathrm{temp}}(q_0)$ for an *arbitrary* parameter $q_0 \in \mathcal{Q}$ by deformation to the generic case. The central support of the discrete series representations of $\mathcal{H}(q_0)$ is given by Theorem 4.9. For a given orbit of residual points $W_0 r$ for $\mathcal{H}(q_0)$ the set of discrete series representations which it carries is, in view of the above, parametrized by the set of orbits of generic residual points $\mathcal{W}_0 r(q)$ such that $W_0 r(q_0) = W_0 r$. This determines the set $\Delta_{F_0}(q_0)$, with complete information about the action of $\mathcal{Z}$. We can repeat this for any standard parabolic subset P, since the associated affine Hecke

algebra $\mathcal{H}_P$ is a tensor product of at most one factor of type C_m^{aff} (with $m \leq n$) (which we can handle as above) and factors of type $A_{\lambda_j - 1}$ (with lattice $X_j = P(A_{\lambda_j - 1})$, the corresponding weight lattice) such that $\sum \lambda_j = n - m$. The group K_P is a product of cyclic groups C_{λ_j} of order λ_j. We also get complete information on the action of $\mathcal{W}$ on Δ : $\mathcal{W}$ is generated by permutations of the type A-factors, by twisting with $-w$ if w is the longest element in the Weyl group of one of the type A-factors, and by twisting of one of the type A factors A_{k-1} by the corresponding cyclic group C_k. Observe that the action of $\mathcal{W}$ only affects the type A-factors of $\mathcal{H}_P$.

From this information we can reconstruct $\mathcal{W}_{\Xi_u}$, and we can verify directly that the cocycle η_Δ (and thus η itself) *is in fact always trivial*. Using Theorem 4.19 we can in principle compute the R-groups (this was actually carried out by Klaas Slooten [63] in the case of real central characters for all $q_0 \in \mathcal{Q}$). By Theorem 5.13 this gives essentially complete information about the structure of the tempered dual $\hat{\mathcal{S}}$.

If $q \in \cap \mathcal{Q}_r$ (intersection over all $r \in \text{Res}$) it is not hard to show that the R-groups are all trivial. This implies that $\mathcal{S}(q)$, in view of Theorem 5.13 and Theorem 5.11, is Morita equivalent to the algebra of $\mathcal{W}$-invariant C^∞-functions on $\Xi_u(q)$. We see that the components of $|\Xi_u(q)|$ are parametrized by the set of ordered triples of partitions (λ, μ, ν) of total weight n. If $\lambda = (1^{m_1}, 2^{m_2}, \ldots, k^{m_k})$ then the component corresponding to (λ, μ, ν) is the product of the quotients $W_0(B_{m_i})\backslash(S^1)^{m_i}$ $(i = 1, \ldots, k)$, and thus homeomorphic to $[0, 1]^{|\lambda|}$. Hence $\mathrm{K}_1(\mathcal{S}(q)) = 0$, and $\mathrm{K}_0(\mathcal{S}(q))$ is the free abelian group generated by the above set of components.

Let us compare this result with the other extreme case $q = 1$. Now $\mathcal{S}(1) = \mathcal{S}_W \simeq C^\infty(T_u) \rtimes W_0$ and thus $\mathrm{K}_*(\mathcal{S}(1)) \simeq \mathrm{K}^*_{W_0}(T_u)$. Application of the equivariant Chern character [7] to this last group yields an isomorphism (after killing torsion) with the periodized cohomology of the extended quotient $W_0\backslash\widehat{T_u}$ of T_u with respect to the action of $W_0 = W(B_n)$. The extended quotient $W_0\backslash\widehat{T_u}$ can be computed directly in this case. It turns out that this space is actually homeomorphic to the orbit space $|\Xi_u(q)|$ which we have just computed in the case where $q \in \cap \mathcal{Q}_r$. This is in complete accordance with Conjecture 2b (but of course, we already used the "discrete part" of this conjecture in order to conclude that each generic residual orbit carries precisely one discrete series representation).

Conjecture 2c gives in this example complete information on the classification problem of the irreducible tempered representations, but not on the internal structure of these representations. In the case of type C_n^{aff} Klaas Slooten [62] defined (for real central character) a "generalized Springer correspondence" in terms of certain symbols (in the sense of Malle [45]) and conjectured that the restriction of these tempered representations to $\mathcal{H}(W_0)$ is precisely given by the generalized Green functions [40], [60] attached to these symbols. These conjectures were verified for $n = 3, 4$.

References

[1] Arthur, J., Eisenstein series and the trace formula. *Proc. Sympos. Pure Math.* **33** (1979), 253–254.

[2] Arthur, J., A Paley-Wiener theorem for real reductive groups. *Acta. Math.* **150** (1983), 1–89.

[3] Arthur, J., On elliptic tempered characters, *Acta. Math.* **171** (1993), 73–138.

[4] Aubert, A.-M., Baum, P. F., Plymen, R. J., The Hecke algebra of a reductive p-adic group: a geometric conjecture. Preprint, 2005; math.RT/0502234.

[5] Barbasch, D., Moy, A., A unitarity criterion for p-adic groups. *Invent. Math.* **98** (1) (1989), 19–37.

[6] Barbasch, D., Moy, A., Reduction to real infinitesimal character in affine Hecke algebras. *J. Amer. Math. Soc.* **6** (3) (1993), 611–630.

[7] Baum, P. F., Connes, A., Chern character for discrete groups. In *A fete of topology*, Academic Press, Boston, MA, 1988, 163–232.

[8] Baum, P. F., Connes, A., Higson, N., Classifying space for proper actions and K-theory of group C^*-algebras. In *C^*-algebras: 1943–1993*, Contemp. Math. 167, Amer. Math. Soc., Providence, RI, 1994, 240–291.

[9] Baum, P. F., Higson, N., Plymen, R. J., Representation theory of p-adic groups: a view from operator algebras. In *The mathematical legacy of Harish-Chandra*, Proc. Sympos. Pure. Math. 68, Amer. Math. Soc., Providence, RI, 2000, 111–149.

[10] Baum, P. F., Nistor, V., Periodic cyclic homology of Iwahori-Hecke algebras. *K-Theory* **27** (4) (2002), 329–357.

[11] Bernstein, J. N., Le "centre" de Bernstein (ed. by P. Deligne). In *Representations of reductive groups over a local field*, Travaux en Cours, Hermann, Paris 1984, 1–32.

[12] Bernstein, J., Deligne, P., Kazhdan, D., Trace Paley-Wiener theorem for reductive p-adic groups. *J. Analyse Math.* **47** (1986), 180–192.

[13] Bernstein, J., Zelevinski, V., Induced representations on reductive p-adic groups. *Ann. Sci. École Norm. Sup.* **10** (1977), 441—-472.

[14] Borel, A., Admissible representations of a semisimple group over a local field with vectors fixed under an Iwahori subgroup. *Invent. Math.* **35** (1976), 233–259.

[15] Bushnell, C. J., Henniart, G., Kutzko, P. C., Towards an explicit Plancherel formula for p-adic reductive groups. Preprint, 2005.

[16] Bushnell, C. J., Kutzko, P. C., Smooth representations of reductive p-adic groups: structure theory via types. *Proc. London Math. Soc.* **77** (3) (1998), 582–634.

[17] Bushnell, C.J., Kutzko, P.C., Types in reductive p-adic groups: the Hecke algebra of a cover. *Proc. Amer. Math. Soc.* **129** (2) (2001), 601–607.

[18] Cherednik, I. V., Double affine Hecke algebras and Macdonald's conjectures. *Ann. of Math.* **141** (1995), 191–216.

[19] Cherednik, I. V., *Double affine Hecke algebras*. London Math. Soc. Lecture Note Ser. 319, Cambridge University Press, Cambridge 2005.

[20] Cuntz, J., Bivariante K-Theorie für lokalkonvexe Algebren und der Chern-Connes-Charakter. *Doc. Math.* **2** (1997), 139–182.

[21] Dat, J.-F., On the K_0 of a p-adic group. *Invent. Math.* **140** (2000), 171–226.

[22] Delorme, P., Opdam, E. M., The Schwartz algebra of an affine Hecke algebra. Preprint, 2003; math.RT/0312517.

[23] Delorme, P., Opdam, E. M., Analytic R-groups of affine Hecke algebras. Preprint, to appear.

[24] Emsiz, E., Opdam, E. M., Stokman, J. V., Periodic integrable systems with delta-potentials. *Comm. Math. Phys.* **264** (1) (2006), 191–225.

[25] Harish-Chandra, Harmonic analysis on real reductive groups. II: Wave packets in the Schwartz space. *Invent. Math.* **36** (1976), 1–55.

[26] Harish-Chandra, Harmonic analysis on real reductive groups. III: The Maass Selberg relations and the Plancherel formula. *Ann. of Math.* **104** (1976), 117–201.

[27] Heckman, G. J., Opdam, E. M., Yang's system of particles and Hecke algebras. *Ann. of Math.* **145** (1997), 139–173.

[28] Heckman, G. J., Opdam, E. M., Harmonic analysis for affine Hecke algebras. In *Current Developments in Mathematics, 1996* (ed. by S.-T. Yau et al.), International Press, Boston, MA, 1997, 37–60.

[29] Heckman, G. J., Schlichtkrull, H., *Harmonic Analysis and Special Functions on Symmetric Spaces*. Perspect. Math. 16, Academic Press, San Diego, CA, 1994.

[30] Heiermann, V., Décomposition spectrale et représentations spéciales d'un groupe réductif p-adique. *J. Inst. Math. Jussieu* **3** (3) (2004), 327–395.

[31] Howlett, R. B., Lehrer, G. I., Induced cuspidal representations and generalized Hecke rings. *Invent. Math.* **58**, (1980), 37–64.

[32] Iwahori, N., Matsumoto, H., On some Bruhat decompositions and the structure of the Hecke rings of p-adic Chevalley groups. *Inst. Hautes Études Sci. Publ. Math.* **25** (1965), 5–48.

[33] Kato, S., Duality for Representations of a Hecke Algebra. *Proc. Amer. Math. Soc.* **119** (3) (1993), 941–946.

[34] Kazhdan, D., Lusztig, G., Proof of the Deligne-Langlands conjecture for affine Hecke algebras. *Invent. Math.* **87** (1987), 153–215.

[35] Kazhdan, D., Nistor, V., Schneider, P., Hochschild and cyclic homology of finite type algebras. *Selecta Math.* (*N.S.*) **4** (2) (1998), 321–359.

[36] Knapp, A. W., Stein, E. M., Intertwining operators for semisimple groups II. *Invent. Math.* **60** (1) (1980), 9–84.

[37] Labesse, J.-P., Pseudo-coefficients très cuspidaux et K-théorie. *Math. Ann.* **291** (1991), 607–616.

[38] Langlands, R. P., *On the functional equations satisfied by Eisenstein series*. Lecture Notes in Math. 544, Springer-Verlag, Berlin 1976.

[39] Lusztig, G., Affine Hecke algebras and their graded version. *J. Amer. Math. Soc.* **2** (3) (1989), 599–635.

[40] Lusztig, G., Green functions and character sheaves. *Ann. of Math.* **131** (1990), 355–408.

[41] Lusztig, G., Classification of unipotent representations of simple p-adic groups. *Internat. Math. Res. Notices* **1995** (11) (1995), 517–589.

[42] Lusztig, G., Cuspidal local systems and graded Hecke algebras III. *Represent. Theory* **6** (2002), 202–242.

[43] Lusztig, G., *Hecke algebras with unequal parameters*. CRM Monogr. Ser. 18, Amer. Math. Soc., Providence, RI, 2003.

[44] Malle, G., Unipotente Grade imprimitiver komplexer Spiegelungsgruppen. *J. Algebra* **177** (1995), 768—-826.

[45] Macdonald, I. G., *Spherical functions on a group of p-adic type*. Publ. Ramanujan Institute 2, Ramanujan Institute, Centre for Advanced Study in Mathematics, University of Madras, Madras 1971.

[46] Macdonald, I. G., *Affine Hecke algebras and orthogonal polynomials.* Cambridge Tracts in Math. 157, Cambridge University Press, Cambridge 2003.

[47] Matsumoto, H., *Analyse harmonique dans les systèmes de Tits bornologiques de type affine*. Lecture Notes in Math. 590, Springer-Verlag, Berlin 1977.

[48] Meyer, R., Homological algebra for Schwartz algebras of reductive p-adic groups. Preprint, 2005; math.RT/0501548.

[49] Moeglin, C., Waldspurger, J.-L, *Spectral decomposition and Eisenstein series*. Cambridge Tracts in Math. 113, Cambridge University Press, Cambridge 1995.

[50] Moerdijk, I., Orbifolds as groupoids: an introduction. In *Orbifolds in mathematics and physics* (Madison, WI, 2001), Contemp. Math. 310, Amer. Math. Soc., Providence, RI, 2002, 205—222.

[51] Morris, L., Tamely ramified intertwining algebras. *Invent. Math.* **114** (1993), 233–274.

[52] Morris, L., Level zero G-types. *Compositio Math.* **118** (2) (1999), 135–157.

[53] Moy, A., Prasad, G., Unrefined minimal K-types for p-adic groups. *Invent. Math.* **116** (1–3) (1994), 393–408.

[54] Opdam, E. M., *Lectures on Dunkl operators for real and complex reflection groups.* MSJ Memoirs 8, Mathematical Society of Japan, Tokyo 2001.

[55] Opdam, E. M., A generating formula for the trace of the Iwahori-Hecke algebra. In *Studies in memory of Issai Schur*, Progr. Math. 210, Birkhäuser, Boston, MA, 2003, 301–323.

[56] Opdam, E. M., On the spectral decomposition of affine Hecke algebras. *J. Inst. Math. Jussieu* **3**(4) (2004), 531–648.

[57] Reeder, M., Formal degrees and L-packets of unipotent discrete series of exceptional p-adic groups (with an appendix by Frank Lübeck). *J. Reine Angew. Math.* **520** (2000), 37–93.

[58] Reeder, M., Euler-Poincaré pairings and elliptic representations of Weyl groups and p-adic groups. *Compositio Math.* **129** (2) (2001), 149–181.

[59] Schneider, P., Stuhler, U., Representation theory and sheaves on the Bruhat-Tits building. *Inst. Hautes Études Sci. Publ. Math.* **85** (1997), 97–191.

[60] Shoji, T., Green polynomials associated to complex reflection groups. *J. Algebra* **245** (2001), 650–694.

[61] Silberger, A. J., The Knapp-Stein dimension theorem for p-adic groups. *Proc. Amer. Math. Soc.* **68** (2) (1978), 243–246; Correction ibid. **76** (1) (1979), 169–170.

[62] Slooten, K., A combinatorial generalization of the Springer correspondence for type B_n. *Acta Appl. Math.* **86** (1–2) (2005), 159–177.

[63] Slooten, K., Reducibility of induced discrete series representations for affine Hecke algebras of type B. Preprint, 2005; math.RT/0511206.

[64] Solleveld, M., Some Fréchet algebras for which the Chern character is an isomorphism. *K-theory*, to appear; math.KT/0505282, 2005.

[65] Vignéras, M.-F., On formal dimensions for reductive p-adic groups. In *Festschrift in honor of I. I. Piatetski-Shapiro on the occasion of his sixtieth birthday*, Part I, Israel Math. Conf. Proc. 2, Weizmann Science Press of Israel, Jerusalem 1990, 225–265.

[66] Waldspurger, J.-L., La formule de Plancherel pour les groupes p-adiques (d'après Harish-Chandra). *J. Inst. Math. Jussieu* **2** (2003), 235–333.

[67] Wassermann, A. J., Cyclic cohomology of algebras of smooth functions on orbifolds. In *Operator algebras and applications I*, London Math. Soc. Lecture Note Ser. 135, Cambridge University Press, Cambridge 1988, 229–244.

Korteweg-de Vries Instituut voor Wiskunde, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

E-mail: opdam@science.uva.nl

Continuous representation theory of p-adic Lie groups

Peter Schneider

Abstract. In this paper we give an overview over the basic features of the continuous representation theory of p-adic Lie groups as it has emerged during the last five years. The main motivation for developing such a theory is a possible extension of the local Langlands program to p-adic Galois representations. This is still very much in its infancy. But in the last section we will describe a first approximation to an extended Langlands functoriality principle for crystalline Galois representations.

Mathematics Subject Classification (2000). Primary 11S37, 22E50; Secondary 11F70.

Keywords. Lie groups, representation theory, Langlands functoriality.

1. Motivation

Throughout the paper we fix a finite extension $L/\mathbb{Q}_p$ and let q denote the cardinality of its residue field. One object of major interest is the absolute Galois group $\mathcal{G}_L := \mathrm{Gal}(\bar{L}/L)$ where $\bar{L}/L$ is an algebraic closure. One obtains a first very coarse idea of its structure by looking at the tower of fields $L \subseteq L^{\mathrm{nr}} \subseteq L^{\mathrm{tr}} \subseteq \bar{L}$ where L^{nr}, resp. L^{tr}, denotes the maximal unramified, resp. tamely ramified, extension of L. Correspondingly we have the subgroups

$$\mathcal{P}_L := \mathrm{Gal}(\bar{L}/L^{\mathrm{tr}}) \subseteq \mathcal{I}_L := \mathrm{Gal}(\bar{L}/L^{\mathrm{nr}}) \subseteq \mathcal{G}_L.$$

All the complication of $\mathcal{G}_L$ is contained in the pro-p-group $\mathcal{P}_L$. The quotient $\mathcal{I}_L/\mathcal{P}_L$ is pro-cyclic of a pro-order prime to p. The quotient $\mathcal{G}_L/\mathcal{I}_L$ even has the natural Frobenius generator ϕ which makes $\mathcal{G}_L/\mathcal{I}_L$ naturally isomorphic to $\hat{\mathbb{Z}}$. In particular, this allows to introduce the Weil group

$$\mathcal{W}_L := \{g \in \mathcal{G}_L : g \equiv \phi^{\alpha(g)} \mathrm{mod}\ \mathcal{I}_L \text{ for some } \alpha(g) \in \mathbb{Z}\};$$

it is topologized by declaring the inertia subgroup $\mathcal{I}_L$ to be open.

Throughout the paper K will denote the coefficient field of whatever kind of representation we like to consider. It always will have characteristic zero and, most of the time, will carry a topology. This introductory section is about representations of the Weil group $\mathcal{W}_L$ by which we mean a continuous homomorphism

$$\rho\colon \mathcal{W}_L \longrightarrow \mathrm{GL}(E)$$

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

where E is a finite dimensional K-vector space. In the case that K is an abstract field, i.e., that $\mathrm{GL}(E)$ carries the discrete topology, we will speak of a smooth representation. Of course, then $\rho(\mathcal{I}_L)$ is a finite group. Even if we take $K = \mathbb{C}$ to be the field of complex numbers with its natural topology we still have, since $\mathcal{I}_L$ is totally disconnected, that $\rho(\mathcal{I}_L)$ is a finite group.

The setting becomes richer if we choose a prime number ℓ different from p and take for $K = \overline{\mathbb{Q}}_\ell$ an algebraic closure of $\mathbb{Q}_\ell$ with its ℓ-adic topology. Then $\mathrm{GL}(E)$ contains an open subgroup which is pro-ℓ. This means that only the image $\rho(\mathcal{P}_L)$ is finite. Nevertheless this setting still can be made "smooth" in the following way. By abuse of language we mean by a Weil–Deligne group representation a pair (ρ', N) consisting of a smooth representation $\rho' \colon \mathcal{W}_L \longrightarrow \mathrm{GL}(E)$ in a finite dimensional K-vector space E and a (nilpotent) endomorphism $N \colon E \longrightarrow E$ such that

$$g \circ N \circ g^{-1} = q^{\alpha(g)} \cdot N \quad \text{for any } g \in \mathcal{W}_L.$$

The point is that by deriving the action (via ρ) on E of the group $\mathcal{I}_L / \ker(\rho)$ one obtains an endomorphism N of E and that dividing ρ through the exponential of N in an appropriate way results in a smooth representation ρ'. In fact, by Grothendieck's abstract monodromy theorem, this construction sets up a natural bijection

$$\begin{array}{l}\text{isomorphism classes of continuous}\\ \text{representations } \rho \colon \mathcal{W}_L \to \mathrm{GL}(E)\end{array} \quad \longleftrightarrow \quad \begin{array}{l}\text{isomorphism classes of Weil–Deligne}\\ \text{group representations } (\rho', N).\end{array}$$

For a more detailed description we refer to [28]. The representation ρ is called Frobenius semisimple if the lifts in $\mathcal{W}_L$ of the Frobenius ϕ act semisimply on E. For the corresponding Weil–Deligne group representation (ρ', N) this amounts to the semisimplicity of the smooth representation ρ'.

In order to state the local Langlands correspondence we let G be the group of L-rational points of some connected reductive group over L. It naturally is a locally compact and totally disconnected group. Similarly as before a representation of G in a K-vector space V (but which here usually will be infinite dimensional) is called smooth if the corresponding map $G \times V \to V$ is continuous with respect to the discrete topology on V. The local Langlands correspondence asserts ([15], [16]) the existence of a distinguished bijection

$$\begin{array}{l}\text{isomorphism classes of } n\text{-dimensional}\\ \text{Frobenius semisimple continuous}\\ \text{representations } \rho \colon \mathcal{W}_L \longrightarrow \mathrm{GL}(E)\end{array} \quad \longleftrightarrow \quad \begin{array}{l}\text{isomorphism classes of irreducible}\\ \text{smooth representation of } \mathrm{GL}_n(L)\end{array}$$

for any $n \geq 1$ (and where the coefficient field still is $K = \overline{\mathbb{Q}}_\ell$). In fact, there is a much more general but conjectural local Langlands functoriality principle. The Langlands dual group ${}^L G$ of G is a semidirect product

$${}^L G = {}^L G^\circ \rtimes \mathcal{G}_L$$

where ${}^L G^\circ$ is the group of K-rational points of the connected reductive group over K whose root datum is dual to the root datum of G (over $\bar{L}$). The functoriality principle asserts that the set $\Pi(G)$ of isomorphism classes of irreducible smooth representations of G has a distinguished partitioning into finite subsets $\Pi_{\rho'}$ which are indexed by equivalence classes of certain Weil–Deligne group representations ρ' with values in ${}^L G$.

There is a particularly simple special case of this functoriality principle which is the unramified correspondence. An irreducible smooth representation of G is called unramified if it has a nonzero vector fixed by a good maximal compact subgroup. Suppose for simplicity that G is L-split. The Satake isomorphism for the Hecke algebra of this maximal compact subgroup produces from an unramified representation of G an orbit of unramified characters of a maximal split torus T in G and hence an orbit of points in ${}^L T^\circ(K)$ where ${}^L T^\circ \subseteq {}^L G^\circ$ is the dual torus. This means one has a natural bijection

$$\begin{array}{l}\text{isomorphism classes of unramified irreducible smooth representations of } G\end{array} \quad \longleftrightarrow \quad \begin{array}{l}\text{semisimple conjugacy classes in } {}^L G^\circ.\end{array}$$

Moreover the semisimple conjugacy class of an $s \in {}^L G^\circ$ corresponds to the equivalence class of the unramified Weil–Deligne group representation $\mathcal{W}_L \longrightarrow \mathcal{W}_L / \mathcal{I}_L \longrightarrow {}^L G^\circ$ which sends the Frobenius ϕ to s. For more details about the functoriality conjecture we refer to [3].

In this paper we are interested in the case $K = \overline{\mathbb{Q}}_p$. Then there is no obvious restriction on the image $\rho(\mathcal{P}_L)$ any more. This means that the p-adic representation theory of $\mathcal{W}_L$ is drastically more complicated than the previous ℓ-adic one. There is a natural functor, constructed by Fontaine ([13]), which associates with any continuous representation $\rho\colon \mathcal{W}_L \longrightarrow \mathrm{GL}(E)$ a Weil–Deligne group representation $Fon(\rho)$ in a free $\hat{\mathbb{Q}}_p^{\mathrm{nr}} \otimes_{\mathbb{Q}_p} K$-module of finite rank; here $\hat{\mathbb{Q}}_p^{\mathrm{nr}}$ denotes the completion of $\mathbb{Q}_p^{\mathrm{nr}}$. It should be viewed as a replacement for the monodromy theorem in the ℓ-adic case. But its construction is much more involved. Moreover, for most ρ one in fact has $Fon(\rho) = 0$. The rich theory of p-adic Galois representations which has evolved during the last twenty years therefore is largely restricted to the so called potentially semistable ones, i.e., to those for which ρ and $Fon(\rho)$ have the same rank. The author nevertheless is very much convinced that there is an extension of the Langlands functoriality principle which takes into account all p-adic representations of $\mathcal{W}_L$. Of course, it should be compatible with the traditional functoriality conjecture via the functor *Fon*.

Since the category of p-adic representations of $\mathcal{W}_L$ is so much richer and more complicated than the category of ℓ-adic representations it is clear that the author's conviction only has a chance if also on the side of the reductive group G we will be able to introduce a much richer but still reasonable category of representations of G than the category of smooth representations. The purpose of this paper is to report on such a construction which was developed in joint work with J. Teitelbaum. At the

end we will come back to Galois representations and will briefly discuss a possible extension of the unramified correspondence.

2. Banach space representations

From now on we always assume K to be a finite extension of $\mathbb{Q}_p$ and we let o denote the ring of integers in K. At first we let G be an arbitrary locally compact and totally disconnected group. Smoothness for a linear G-action on a K-vector space means continuity with respect to the discrete topology on the vector space. It therefore is a rather obvious idea to enrich the picture by considering continuous linear G-actions on some class of topological K-vector spaces.

A reasonable framework for topological K-vector spaces is provided by the notion of a locally convex topology. Such topologies on a K-vector space are defined by a family of nonarchimedean seminorms (cf. [21], §4). The open convex neighborhoods of the zero vector in a locally convex K-vector space are lattices, i.e. o-submodules which generate the vector space.

The most straightforward class of locally convex K-vector spaces is the class of K-Banach spaces. A K-Banach space is complete and its topology can be defined by a single nonarchimedean norm. A K-Banach space representation of G is a linear G-action on a K-Banach space V such that the corresponding map $G \times V \to V$ is continuous. Unfortunately this definition is much too general in order to lead to a reasonable category. We mention only two pathological phenomena:

- There can exist nonzero G-equivariant continuous linear maps between two nonisomorphic topologically irreducible Banach space representations.

- Already such a simple group as the additive group of p-adic integers $G = \mathbb{Z}_p$ has infinite dimensional topologically irreducible Banach space representations ([8]).

The challenge is to impose an additional "finiteness" condition leading to a category of representations which at the same time is rich enough and still is manageable.

To prepare the subsequent definition taken from [23] we point out the following. Let V be any K-Banach space representation of G. Given any compact open subgroup $H \subseteq G$ and any open lattice $M \subseteq V$ the o-submodule $\cap_{g \in H} gM$ is an H-invariant open lattice in V. We also recall that an o-module N is called of cofinite type if its Pontrjagin dual $\mathrm{Hom}_o(N, K/o)$ is a finitely generated o-module.

Definition 2.1. A K-Banach space representation V of G is called admissible if for any compact open subgroup $H \subseteq G$, for any bounded H-invariant open lattice $M \subseteq V$, and for any open subgroup $H' \subseteq H$ the o-submodule $(V/M)^{H'}$ of H'-invariant elements in the quotient V/M is of cofinite type.

We let $\mathrm{Ban}^a_G(K)$ denote the category of all admissible K-Banach space representations of G with continuous linear G-equivariant maps.

A first justification for this definition might be the following observation. We recall that a smooth representation of G is called admissible if the subspace of H'-fixed vectors, for any compact open subgroup $H' \subseteq G$, is finite dimensional. Let V be a K-Banach space representation of G, let $H \subseteq G$ be a compact open subgroup, and let $M \subseteq V$ be a bounded H-invariant open lattice. Then $M/\pi M$, where π is a prime element in o, evidently is a smooth representation of H over the residue field of K. If V is admissible then $M/\pi M$ is an admissible smooth representation of H.

In which sense is this category $\mathrm{Ban}^a_G(K)$ manageable? In order to answer this question we have to assume that G is a locally $\mathbb{Q}_p$-analytic group, i.e., a p-adic Lie group. Moreover, since the definition of admissibility is in terms of compact open subgroups it suffices to discuss this issue for compact groups. In the following we therefore let H be a compact p-adic Lie group. The completed group ring of H is defined to be

$$o[[H]] := \varprojlim o[H/H']$$

where H' runs over all open normal subgroups of H. This is a compact linear-topological o-algebra. The H-action on any K-Banach space representation extends uniquely to a separately continuous $o[[H]]$-module structure. We have the following crucial fact due to Lazard ([18], V.2.2.4).

Theorem 2.2. *For any compact p-adic Lie group H the ring $o[[H]]$ is noetherian.*

To understand how we will explore this fact let us look at the vector space $C(H, K)$ of all K-valued continuous functions on H. This is a K-Banach space representation of H for the sup-norm and the left translation action of H. But it is difficult to say anything straightforward about the corresponding $o[[H]]$-module $C(H, K)$. Instead, let us pass to the continuous dual

$$D^c(H, K) := C(H, K)'.$$

First of all $D^c(H, K)$ is a K-algebra with respect to the convolution product of continuous linear forms – the algebra of continuous distributions on H. Secondly, sending an element $g \in H$ to the Dirac distribution $\delta_g \in D^c(H, K)$ extends to an embedding of o-algebras

$$o[[H]] \hookrightarrow D^c(H, K)$$

whose image is a lattice so that, in fact, $K \otimes_o o[[H]] \cong D^c(H, K)$. Thirdly, via the latter isomorphism, the $K \otimes_o o[[H]]$-module structure on $C(H, K)'$ induced by functoriality from the $o[[H]]$-module structure on $C(H, K)$ simply corresponds to the action of $D^c(H, K)$ on itself by multiplication. Quite generally, the continuous dual V' of a K-Banach space representation V of H, by functoriality, is a module over the K-algebra $D^c(H, K)$. The main result (Thm. 3.5) of [23] now is the following.

Theorem 2.3. *Let* $\mathrm{Mod}_{fg}(D^c(H,K))$ *denote the category of finitely generated* $D^c(H,K)$*-modules. The functor*

$$\begin{aligned}\mathrm{Ban}^a_H(K) &\xrightarrow{\sim} \mathrm{Mod}_{fg}(D^c(H,K)),\\ V &\longmapsto V'\end{aligned}$$

is an anti-equivalence of categories.

By Theorem 2.2 the ring $D^c(H,K)$ is noetherian. Hence with $\mathrm{Mod}_{fg}(D^c(H,K))$ also $\mathrm{Ban}^a_G(K)$ is an abelian category. We mention that the underlying vector space of the kernel, image, and cokernel of a morphism in $\mathrm{Ban}^a_G(K)$ is the kernel, image, and cokernel, respectively, of the underlying linear map. We also emphasize that the above result completely algebraizes the theory of admissible K-Banach space representations.

As evidence for the category $\mathrm{Ban}^a_G(K)$ being rich enough we will discuss the continuous principal series of the group G of L-rational points of a connected reductive group over L. Let $P \subseteq G$ be a parabolic subgroup and $\chi : P \to K^\times$ be a continuous character. We put

$${}^c\mathrm{Ind}_P^G(\chi) := \{f : G \to K \text{ continuous: } f(gb) = \chi(b)^{-1} f(g) \text{ for } g \in G,\ b \in P\}$$

on which G acts by left translations. If $G_0 \subseteq G$ is a good maximal compact subgroup then we have the Iwasawa decomposition $G = G_0 P$. Hence taking the supremum over G_0 defines a norm on ${}^c\mathrm{Ind}_P^G(\chi)$. A different choice of G_0 leads to an equivalent norm.

Proposition 2.4. ${}^c\mathrm{Ind}_P^G(\chi)$ *is an admissible* K*-Banach space representation of the reductive group* G*.*

Proof. We only indicate the argument for the admissibility. Restricting functions to G_0 defines a topological embedding ${}^c\mathrm{Ind}_P^G(\chi) \to C(G_0, K)$. Dually this exhibits the continuous dual ${}^c\mathrm{Ind}_P^G(\chi)'$ as a quotient of $D^c(G_0, K)$. □

It seems likely that the representation ${}^c\mathrm{Ind}_P^G(\chi)$ always has a finite composition series. We want to state a precise irreducibility conjecture. To avoid technicalities we assume that $L \subseteq K$, that G is L-split, semisimple, and simply connected, and that P is a Borel subgroup. Let $T \subseteq P$ be a maximal L-split torus. Then χ can be viewed as a continuous character $\chi : T \to K^\times$. Let $X^*(T)$ resp. $X_*(T)$, denote as usual the group of rational characters, resp. cocharacters, of T. In $X^*(T)$ we have the subsets $\Delta \subseteq \Phi^+$ of simple and of positive roots with respect to P, respectively. For any $\alpha \in \Phi^+$ there is the corresponding coroot $\check{\alpha} \in X_*(T)$. By our assumption on G any fundamental weight ω_α, for $\alpha \in \Delta$, and hence their sum $\delta := \sum_{\alpha\in\Delta} \omega_\alpha$ lie in $X^*(T)$. The character $\chi : T \to K^\times$ is called anti-dominant if $\chi\delta \circ \check{\alpha} \neq (\ )^m$ for any integer $m \geq 1$ and any $\alpha \in \Phi^+$. Here $(\ )^m : L^\times \to K^\times$, for any $m \in \mathbb{Z}$, is the continuous character sending a to a^m.

Conjecture 2.5. The G-representation ${}^c\mathrm{Ind}_P^G(\chi)$ is topologically irreducible if χ^{-1} is anti-dominant.

Proposition 2.6. *Suppose that $L = \mathbb{Q}_p$. We then have:*

i. *The above conjecture holds true for the group $G = \mathrm{GL}_2(\mathbb{Q}_p)$.*

ii. *If the anti-dominance condition for χ^{-1} continues to hold after restriction to an arbitrary small open subgroup of $\mathbb{Q}_p^\times$ then the continuous dual ${}^c\mathrm{Ind}_P^G(\chi)'$ is simple as a $D^c(G_0, K)$-module and ${}^c\mathrm{Ind}_P^G(\chi)$, in particular, is topologically irreducible as a G_0-representation.*

The proof of this result is highly indirect. It requires corresponding facts for the locally analytic principal series (see Section 3) and the density of analytic vectors in admissible Banach space representations (see Section 4).

Breuil has introduced the notion of a unitary Banach space representation of G which means that the Banach space topology can be defined by a G-invariant norm. In recent work Berger and Breuil ([2]) and Colmez ([5]) construct, by very sophisticated methods, a series of topologically irreducible, admissible, and unitary Banach space representations of the group $\mathrm{GL}_2(\mathbb{Q}_p)$. In fact, they put this series into correspondence with certain two dimensional Galois representations of $\mathcal{G}_{\mathbb{Q}_p}$ called trianguline by Colmez.

We also point out that in a unitary Banach space representation V of G we find a G-invariant open lattice $M \subseteq V$. If V is admissible then $M/\pi M$, with $\pi \in o$ a prime element, is an admissible smooth representation of G over the residue field of K. As to be expected, the theory described in this section therefore is closely related to the smooth representation theory of p-adic groups with characteristic p coefficients (cf. [29]).

3. Locally analytic representation

Many continuous representations of p-adic Lie groups in locally convex K-vector spaces which naturally arise in geometric situations are not Banach space representations. One basic example is the following. Let $G = \mathrm{GL}_2(L)$ and let $\mathcal{X}_L := \mathbb{P}^1(\bar{L})\backslash\mathbb{P}^1(L)$ be the p-adic upper half plane over L. The rigid analytic variety $\mathcal{X}_L$ carries an obvious G-action. The vector space $\Omega^1(\mathcal{X}_L)$ of global holomorphic 1-forms on $\mathcal{X}_L$ is an L-Fréchet space with a natural continuous G-action. By a theorem of Morita ([19]) its continuous dual is isomorphic to the quotient $C^{\mathrm{an}}(\mathbb{P}^1(L), L)/L$ of the vector space $C^{\mathrm{an}}(\mathbb{P}^1(L), L)$ of L-valued locally L-analytic functions on the projective line $\mathbb{P}^1(L)$ by the subspace of constant functions. Clearly, $C^{\mathrm{an}}(\mathbb{P}^1(L), L)$ is not a Banach space. In fact, its natural locally convex topology is rather complicated. But in return it has good properties like being reflexive.

This observation leads to the concept of a locally analytic representation of a locally L-analytic group G. For the rest of the paper we assume that $L \subseteq K$. We first

remark that for any paracompact locally L-analytic manifold X and any Hausdorff locally convex K-vector space V the K-vector space $C^{\mathrm{an}}(X, V)$ of V-valued locally analytic functions on X is well defined. It carries a natural Hausdorff locally convex topology for whose rather technical construction we refer to [12].

Definition 3.1. A locally analytic representation V of G (over K) is a barrelled locally convex Hausdorff K-vector space V equipped with a G-action by continuous linear endomorphisms such that, for each $v \in V$, the map $g \mapsto gv$ lies in $C^{\mathrm{an}}(G, V)$.

The requirement that V is barrelled (i.e., that each closed lattice in V is open) is a convenient (and mild) technical restriction which makes applicable the Banach–Steinhaus theorem ([21], Prop. 6.15). It implies that the map $G \times V \to V$ describing the G-action is continuous and that the Lie algebra $\mathfrak{g}$ of G acts continuously on V by

$$\begin{aligned} \mathfrak{g} \times V &\longrightarrow V, \\ (\mathfrak{x}, v) &\longmapsto \mathfrak{x}v := \frac{d}{dt} \exp(t\mathfrak{x}) v_{|t=0}. \end{aligned}$$

The next step is, as in the previous section, to pass from the G-action to a module structure. The strong dual

$$D(G, K) := C^{\mathrm{an}}(G, K)'_b$$

is called the locally convex vector space of K-valued locally analytic distributions on G.

Proposition 3.2. i. *The convolution on $D(G, K)$ is well defined, is separately continuous, and makes $D(G, K)$ into an associative K-algebra with the Dirac distribution δ_1 in $1 \in G$ as the unit element.*

ii. *The map*

$$\begin{aligned} \mathfrak{g} &\longrightarrow D(G, K), \\ \mathfrak{x} &\longmapsto [f \mapsto ((-\mathfrak{x})f)(1)] \end{aligned}$$

extends to a monomorphism of K-algebras $U(\mathfrak{g}) \otimes_L K \longrightarrow D(G, K)$ where $U(\mathfrak{g})$ denotes the universal enveloping algebra of $\mathfrak{g}$.

iii. *The G-action on any locally analytic representation V of G extends uniquely to a separately continuous action of the algebra $D(G, K)$ on V.*

Proof. i. [11], 4.4.1 and 4.4.4, or [26], Remark A.1. ii. [24], p. 450. iii. [24], Prop. 3.2. □

It is important to notice ([24], Lemma 2.1) that for a compact group G the locally convex topology on $C^{\mathrm{an}}(G, K)$ is of compact type ([24], §1, and [21], §16). In particular, $D(G, K)$ then is a Fréchet algebra. Locally convex topologies of compact type are rather complicated but have the remarkable property that they can be characterized by their strong dual being nuclear and Fréchet. This leads to the following result ([24], Cor. 3.4).

Proposition 3.3. *If G is compact then the functor*

$$\begin{array}{ll} \textit{locally analytic representations of } G & \textit{continuous } D(G,K)\textit{-modules on} \\ \textit{on } K\textit{-vector spaces of compact type} & \xrightarrow{\sim} \quad \textit{nuclear Fréchet spaces with} \\ \textit{with continuous linear } G\textit{-maps} & \textit{continuous } D(G,K)\textit{-module maps} \end{array}$$

which sends V to its strong dual V_b' is an anti-equivalence of categories.

This last result means that in the context of locally analytic representations we now have singled out the correct type of locally convex topology to which we should restrict our attention. But we are still lacking the equivalent of the algebraic finiteness condition which we have imposed in the last section. The situation is considerably complicated by the fact that even for compact G the algebra $D(G, K)$ is far from being noetherian. As a remedy for this we develop in [25], §3, the following axiomatic framework.

Suppose that A is a K-Fréchet algebra. For any continuous algebra seminorm q on A the completion A_q of A with respect to q is a K-Banach algebra. For any two such seminorms $q' \leq q$ the identity on A extends to a continuous K-algebra homomorphism $\phi_q^{q'} : A_q \to A_{q'}$.

Definition 3.4. A K-Fréchet algebra A is called a Fréchet–Stein algebra if there is a sequence $q_1 \leq \cdots \leq q_n \leq \cdots$ of continuous algebra seminorms on A which define the Fréchet topology and such that

(i) A_{q_n} is (left) noetherian, and

(ii) A_{q_n} is flat as a right $A_{q_{n+1}}$-module (via $\phi_{q_{n+1}}^{q_n}$)

for any $n \in \mathbb{N}$.

Suppose that A is a Fréchet–Stein algebra as in the above definition. We have the obvious isomorphism of Fréchet algebras

$$A \xrightarrow{\cong} \varprojlim_n A_{q_n}.$$

Definition 3.5. A (left) A-module N is called coadmissible if

(i) $A_{q_n} \otimes_A N$ is finitely generated over A_{q_n} for any $n \in \mathbb{N}$, and

(ii) the natural map $N \xrightarrow{\cong} \varprojlim_n (A_{q_n} \otimes_A N)$ is an isomorphism.

We let $\mathcal{C}_A$ denote the full subcategory of all coadmissible modules in the category $\mathrm{Mod}(A)$ of all left A-modules. By a cofinality argument it is independent of the particular choice of the sequence of seminorms q_n.

Proposition 3.6. *$\mathcal{C}_A$ is an abelian subcategory of* $\mathrm{Mod}(A)$ *containing all finitely presented A-modules.*

Over a noetherian Banach algebra every finitely generated module carries a unique Banach space topology which makes the module structure continuous. Using its representation as a projective limit in Definition 3.5 (ii) we therefore see that any coadmissible A-module carries a distinguished Fréchet topology which will be called its canonical topology.

In [25], Thm. 5.1, we give the following justification for introducing these definitions.

Theorem 3.7. *For any compact locally L-analytic group G the Fréchet algebra $D(G, K)$ is a K-Fréchet–Stein algebra.*

If $G_{\mathbb{Q}_p}$ denotes the locally $\mathbb{Q}_p$-analytic group which underlies G then one has a natural continuous surjection

$$D(G_{\mathbb{Q}_p}, K) \longrightarrow D(G, K).$$

By a general argument with Fréchet–Stein algebras ([25], Prop. 3.7) this allows to reduce the proof of the theorem to the case $L = \mathbb{Q}_p$. Furthermore it is easy to see that it suffices to prove the theorem for a conveniently chosen open normal subgroup of G. To avoid technicalities we assume in the following that $p \neq 2$. Let ω_p denote the additive p-adic valuation on $\mathbb{Z}_p$. Lazard in [18], III.2.1.2, has introduced the notion of a p-valuation on a group H which is a real valued function $\omega\colon H\backslash\{1\} \longrightarrow (1/(p-1), \infty)$ such that

$$\omega(gh^{-1}) \geq \min(\omega(g), \omega(h)),$$
$$\omega(g^{-1}h^{-1}gh) \geq \omega(g) + \omega(h),$$

and

$$\omega(g^p) = \omega(g) + 1.$$

Our proof of the above theorem very much relies on the techniques and results in [18]. First of all we may assume, by the above reductions and [9], Cor. 8.34 (ii), that G is a compact p-adic Lie group which has an (ordered) set of topological generators $h_1, \ldots, h_d$ such that:

(i) The map

$$\mathbb{Z}_p^d \xrightarrow{\sim} G,$$
$$(x_1, \ldots, x_d) \longmapsto h_1^{x_1} \ldots h_d^{x_d}$$

is a global chart of the manifold G.

(ii) The function

$$\omega(h_1^{x_1} \ldots h_d^{x_d}) := \min_{1 \leq i \leq d} (1 + \omega_p(x_i))$$

is a p-valuation on G.

(iii) Every $g \in G$ such that $\omega(g) \geq 2$ is a p-th power in G.

We define $b_i := \delta_{h_i} - 1$ and, for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$,

$$\boldsymbol{b}^\alpha := b_1^{\alpha_1} \ldots b_d^{\alpha_d} \in D(G, K)$$

(which does depend on the ordering of the generators h_i!). Using the global chart in (i) we may identify $D(G, K)$ as a Fréchet space with $D(\mathbb{Z}_p^d, K)$. The latter, by Amice's p-adic Fourier isomorphism ([1]), is naturally isomorphic to the ring of power series with coefficients in K converging on the open unit polydisk. It follows that any distribution $\lambda \in D(G, K)$ has a unique convergent expansion

$$\lambda = \sum_{\alpha \in \mathbb{N}_0^d} d_\alpha \boldsymbol{b}^\alpha$$

where the set $\{|d_\alpha| r^{\alpha_1 + \cdots + \alpha_d}\}_{\alpha \in \mathbb{N}_0^d}$, for any $0 < r < 1$, is bounded. Moreover, the family of norms

$$\|\lambda\|_r := \sup_{\alpha \in \mathbb{N}_0^d} |d_\alpha| r^{\alpha_1 + \cdots + \alpha_d}$$

defines the Fréchet topology on $D(G, K)$. Lazard in [18] investigates the norm $\| \ \|_{\frac{1}{p}}$.

Lemma 3.8. *For each $\frac{1}{p} \leq r < 1$ the norm $\| \ \|_r$ is submultiplicative.*

The completion $D_r(G, K)$ of $D(G, K)$ with respect to the norm $\| \ \|_r$ for $\frac{1}{p} \leq r < 1$, is a K-Banach algebra and

$$D(G, K) = \varprojlim_{\frac{1}{p} \leq r < 1} D_r(G, K).$$

Proposition 3.9. *For $\frac{1}{p} < r' \leq r < 1$ in $p^{\mathbb{Q}}$ we have:*

i. *$D_r(G, K)$ is noetherian with multiplicative norm $\| \ \|_r$.*

ii. *The homomorphism $D_r(G, K) \longrightarrow D_{r'}(G, K)$ is flat.*

This is proved in [25], following the model for $\| \ \|_{\frac{1}{p}}$ in [18], by explicitly computing the associated graded ring for the filtration on $D_r(G, K)$ defined by the norm $\| \ \|_r$. Philosophically this technique means to view the noncommutative ring $D_r(G, K)$ as a deformation quantization of the commutative ring $D_r(\mathbb{Z}_p^d, K)$ which allows to transfer many ring theoretic properties from the latter to the former.

For later purposes we mention the following important additional result in [14], §1.4, Cor. 2.

Theorem 3.10 (Frommer). *Suppose that G is as above and that $\frac{1}{p} < r < 1$ in $p^{\mathbb{Q}}$. Then $D_r(G, K)$ is finitely generated free as a module over the closure of the universal enveloping algebra $U(\mathfrak{g}) \otimes_L K$.*

By [17], Thm. 1.4.2, this result continues to hold for $L \neq \mathbb{Q}_p$ but the precise conditions on the compact locally L-analytic group G are a little too technical to be formulated here. Having Theorem 3.7 at our disposal we now make the following definition where G again is a general locally L-analytic group.

Definition 3.11. A locally analytic representation of G on a K-vector space of compact type V is called admissible if its strong dual V_b' as a $D(H, K)$-module, for some (or equivalently any) compact open subgroup $H \subseteq G$, is coadmissible with its canonical topology.

Using Proposition 3.6 we obtain the following facts about the category $\mathrm{Rep}_G^a(K)$ of all admissible locally analytic representations of G ([25], Prop. 6.4).

Proposition 3.12. i. $\mathrm{Rep}_G^a(K)$ *is an abelian category; kernel and image of a morphism in* $\mathrm{Rep}_G^a(K)$ *are the algebraic kernel and image with the subspace topology.*

ii. *Any morphism in* $\mathrm{Rep}_G^a(K)$ *is strict and has closed image.*

iii. *The category* $\mathrm{Rep}_G^a(K)$ *is closed with respect to the passage to closed G-invariant subspaces.*

A first evidence that this category $\mathrm{Rep}_G^a(K)$ is not too small is provided by the following result ([25], Thm. 6.6). Let $\mathrm{Rep}_G^{\infty,a}(K)$ denote the abelian category of admissible smooth representations of G in K-vector spaces as recalled in Section 2.

Theorem 3.13. *The functor* $\mathrm{Rep}_G^{\infty,a}(K) \longrightarrow \mathrm{Rep}_G^a(K)$ *of equipping an admissible smooth representation with the finest locally convex topology is a fully faithful embedding; its image is characterized by the condition that the Lie algebra* $\mathfrak{g}$ *acts trivially* (*i.e.,* $\mathfrak{g}V = 0$).

This says that the representation theory appearing in the Langlands functoriality conjecture is fully contained in the new admissible locally analytic theory. It is rather obvious that in case G is an algebraic group also every rational representation of G is admissible locally analytic. Similarly as in the previous section there is a locally analytic principal series for any connected reductive group G over L. Let $P \subseteq G$ be a parabolic subgroup and $\chi : P \longrightarrow K^\times$ be a locally L-analytic character. Then

$$\mathrm{Ind}_P^G(\chi) := \{f : G \to K \text{ locally analytic: } f(gb) = \chi(b)^{-1} f(g) \text{ for } g \in G,\ b \in P\}$$

with G acting by left translations is an admissible locally analytic representation of G. This is proved quite similarly as in the Banach space case. Again we expect that $\mathrm{Ind}_P^G(\chi)$ always has a finite composition series. But there are important differences. For example, if χ is locally constant then the smooth induction ${}^\infty\mathrm{Ind}_P^G(\chi)$ is a proper closed subspace of $\mathrm{Ind}_P^G(\chi)$ but is dense in ${}^c\mathrm{Ind}_P^G(\chi)$. At the moment the deepest result about the locally analytic principal series is the following. If $\mathfrak{p}$ denotes the Lie algebra of P then we have the derived character $d\chi : \mathfrak{p} \longrightarrow K$ and we may form the Verma module $\mathcal{V}_{d\chi} := U(\mathfrak{g}) \otimes_{U(\mathfrak{p})} K_{d\chi}$ of $U(\mathfrak{g}) \otimes_L K$.

Theorem 3.14 (Frommer). *Suppose that $L = \mathbb{Q}_p$ and that G is $\mathbb{Q}_p$-split; if $\mathcal{V}_{-d\chi}$ is a simple module for $U(\mathfrak{g}) \otimes_L K$ then $\mathrm{Ind}_P^G(\chi)$ is topologically irreducible as a representation of G*

It is his Theorem 3.10 which allows Frommer to make this close connection to Verma modules. In view of Kohlhaase's generalization of Theorem 3.10 it seems very likely that Theorem 3.14 holds for any L-split group G where L is arbitrary. Without recalling the details we mention that the simplicity of Verma modules is decided by anti-dominance properties of the inducing character. In [22], §4, we determine, in the case of the group $G = \mathrm{SL}_2(\mathbb{Q}_p)$, a complete composition series for the reducible locally analytic principal series; its length is two or three.

We want to briefly mention three further results in the locally analytic theory. Extending Harish Chandra's computation of the center of $U(\mathfrak{g})$ Kohlhaase determines in [17] the center of $D(G, K)$. His result is particularly simple to formulate in the following case ([17], Thms. 2.1.6 and 2.4.2).

Theorem 3.15 (Kohlhaase). *Suppose that G is an L-split connected reductive group with maximal L-split torus T and corresponding Weyl group W. There is a canonical isomorphism*

$$\text{center of } D(G, K) \cong D(T, K)_Z^W$$

where the right hand side denotes the subalgebra of $D(T, K)$ consisting of all W-invariant distributions supported on the center Z of G. Furthermore, if in addition $Z = \{1\}$ then $D(T, K)_Z^W$ is isomorphic to the ring of entire functions on the rigid analytic affine space over K of dimension equal to the rank of T.

A very basic construction in representation theory is the construction of the contragredient representation. But in the present context the strong dual V_b' of a K-vector space of compact type V is a Fréchet space and hence rarely is again of compact type. The naive method to define the contragredient representation therefore does not work in the locally analytic theory. Perhaps this is not too surprising since we already have used up, so to speak, the strong dual to define the notion of admissibility. In [26] we construct in case $L = \mathbb{Q}_p$, as a replacement for the contragredient, a natural auto-antiequivalence of a certain derived category of $\mathrm{Rep}_G^a(K)$. This is based on the fact that the rings $D_r(G, K)$, for a p-valued Lie group G as discussed above, are Auslander regular, hence that the category $\mathcal{C}_G$ in an appropriate sense is Auslander regular, and it then uses the derived functor of the functor $\mathrm{Hom}_{D(G,K)}(\,.\,, D(G, K))$ on $\mathcal{C}_G$. The restriction to the case $L = \mathbb{Q}_p$ will be removed in the forthcoming thesis of T. Schmidt.

As exemplified by the various principal series, a very important method to construct representations is, in the case of a reductive group G, the induction from the Levi quotient of a parabolic subgroup. In the smooth theory this functor, called parabolic induction, has an adjoint functor in the opposite direction, called parabolic restriction or Jacquet functor. In [10] Emerton constructs by rather sophisticated techniques a

generalization of the Jacquet functor to locally analytic representations. Although no longer adjoint to parabolic induction it is doubtlessly an important construction which must be investigated further.

4. Analytic vectors

In order to describe the connection between Banach space and locally analytic representations we let V be a K-Banach space representation of the locally L-analytic group G.

Definition 4.1. A vector $v \in V$ is called analytic if the (continuous) map $\rho_v(g) := g^{-1}v$ lies in $C^{\mathrm{an}}(G, V)$.

Obviously $V_{\mathrm{an}} := \{v \in V : v \text{ is analytic}\}$ is a G-invariant subspace of V. But we always equip V_{an} with the subspace topology with respect to the G-equivariant embedding

$$\begin{aligned} V_{\mathrm{an}} &\longrightarrow C^{\mathrm{an}}(G, V), \\ v &\longmapsto \rho_v. \end{aligned}$$

One checks that V_{an} is closed in $C^{\mathrm{an}}(G, V)$. Of course, in this generality the subspace V_{an} might very well be zero. But in [25], Thm. 7.1, the following is shown.

Theorem 4.2. *Suppose that $L = \mathbb{Q}_p$; if V is admissible then V_{an} is dense in V and is an admissible locally analytic representation of G; moreover the functor*

$$\begin{aligned} \mathrm{Ban}^a_G(K) &\longrightarrow \mathrm{Rep}^a_G(K), \\ V &\longmapsto V_{\mathrm{an}} \end{aligned}$$

is exact.

The key reason for this result is the following purely algebraic fact ([25], Thm. 5.2).

Theorem 4.3. *Suppose that $L = \mathbb{Q}_p$ and that G is compact; then the natural ring homomorphism*

$$D^c(G, K) \longrightarrow D(G, K)$$

is faithfully flat.

In the case of the principal series for a locally L-analytic character χ we, of course, have

$${}^c\mathrm{Ind}_P^G(\chi)_{\mathrm{an}} = \mathrm{Ind}_P^G(\chi).$$

It is a remarkable fact that ${}^c\mathrm{Ind}_P^G(\chi)$ and $\mathrm{Ind}_P^G(\chi)$ can have different length. In particular, the former can be topologically irreducible and the latter not.

5. Unramified p-adic functoriality

For the simplicity of the presentation we assume in this section that the base field is $L = \mathbb{Q}_p$. As before $K/\mathbb{Q}_p$ is a finite extension of which we assume that it contains a square root $p^{1/2}$ of p. The group G is assumed to be $\mathbb{Q}_p$-split. We let $T \subseteq G$ be a maximal split torus and we fix a maximal compact subgroup $U \subseteq G$ which is special for T. The Satake–Hecke algebra $\mathcal{H}(G, 1_U)$ is the convolution algebra of K-valued U-bi-invariant compactly supported functions on G. By the Satake isomorphism this algebra is commutative and its characters ζ into $\overline{K}$ are in natural bijection with the semisimple conjugacy classes $s(\zeta)$ in ${}^L G^\circ(\overline{K})$. Representation theoretically the Satake–Hecke algebra can be described as the algebra of endomorphisms of the smooth G-representation $\mathrm{ind}_U^G(1_U)$ over K obtained by compact induction from the trivial representation 1_U of U. Any character ζ of $\mathcal{H}(G, 1_U)$ therefore gives rise, by specialization, to the smooth G-representation

$$\mathrm{ind}_U^G(1_U) \underset{\mathcal{H}(G,1_U)}{\otimes} \overline{K}_\zeta$$

which has a unique irreducible quotient V_ζ. The correspondence $V_\zeta \longleftrightarrow s(\zeta)$ is the unramified (smooth) Langlands functoriality we have alluded to already in the first section. We also repeat that $s(\zeta)$ should be viewed as the Weil–Deligne group representation $\mathcal{W}_{\mathbb{Q}_p} \to \mathcal{W}_{\mathbb{Q}_p}/\mathcal{I}_{\mathbb{Q}_p} \to {}^L G^\circ(\overline{K})$ which sends the Frobenius ϕ to $s(\zeta)$ (and with $N = 0$).

We now broaden the picture by bringing in an irreducible $\mathbb{Q}_p$-rational representation σ of G of highest weight $\xi \in X^*(T)$. The corresponding Satake–Hecke algebra $\mathcal{H}(G, \sigma_U)$ is the convolution algebra over K of all compactly supported functions $\psi : G \to \mathrm{End}_K(\sigma)$ satisfying

$$\psi(u_1 g u_2) = \sigma(u_1) \circ \psi(g) \circ \sigma(u_2) \quad \text{for any } u_1, u_2 \in U \text{ and } g \in G.$$

Again the algebra $\mathcal{H}(G, \sigma_U)$ can naturally be identified with the algebra of endomorphisms of the compact induction $\mathrm{ind}_U^G(\sigma_U)$ of the restriction $\sigma_U := \sigma|U$. In fact, since σ is a representation of the full group G the algebras $\mathcal{H}(G, \sigma_U)$ and $\mathcal{H}(G, 1_U)$ are isomorphic. But fixing once and for all a U-invariant norm $\| \ \|$ on the K-vector space which underlies σ we may equip $\mathcal{H}(G, \sigma_U)$ with the sup-norm

$$\|\psi\|_\xi := \sup_{g \in G} \|\psi(g)\|$$

where on the right hand side $\| \ \|$ refers to the operator norm on $\mathrm{End}_K(\sigma)$. Since $\| \ \|_\xi$ obviously is submultiplicative the algebra $\mathcal{H}(G, \sigma_U)$ gives rise, by completion with respect to $\| \ \|_\xi$, to a K-Banach algebra $\mathcal{B}(G, \sigma_U)$. Clearly, $\mathcal{B}(G, \sigma_U)$ is very far from being isometrically isomorphic to $\mathcal{B}(G, 1_U)$. Correspondingly we have a sup-norm on $\mathrm{ind}_U^G(\sigma_U)$ which by completion leads to a unitary Banach space representation $B_U^G(\sigma_U)$ of G.

Lemma 5.1. *$\mathcal{B}(G, \sigma_U)$ is isometrically isomorphic to the algebra of continuous G-equivariant endomorphisms of the Banach space $B_U^G(\sigma_U)$.*

Proof. [27], Lemma 1.3. □

As the completion of a commutative algebra $\mathcal{B}(G, \sigma_U)$ of course is commutative as well. For any K-valued (continuous) character ζ of $\mathcal{B}(G, \sigma_U)$ we obtain, by specialization, the unitary Banach space representation

$$B_{\xi,\zeta} := B_U^G(\sigma_U) \hat{\otimes}_{\mathcal{B}(G,\sigma_U)} K_\zeta$$

of G (where $\hat{\otimes}$ denotes the completed tensor product). Unfortunately the following conjecture seems to be a very difficult problem.

Conjecture 5.2. $B_{\xi,\zeta}$ always is nonzero.

On the other hand it is not to be expected that the Banach space representations $B_{\xi,\zeta}$ are admissible in general. One of the main results in [27] is the explicit computation of the Banach algebra $\mathcal{B}(G, \sigma_U)$. For this we let $\omega_p \colon K^\times \to \mathbb{R}$ denote the unique additive valuation such that $\omega_p(p) = 1$ and we introduce the map

$$\mathrm{val} : {}^L T^\circ(K) = \mathrm{Hom}(T/U \cap T, K^\times) \xrightarrow{\omega_{p^\circ}} \mathrm{Hom}(T/U \cap T, \mathbb{R}) =: V_{\mathbb{R}}.$$

We note that via the isomorphism

$$\begin{aligned} X^*(T) \otimes \mathbb{R} &\xrightarrow{\cong} V_{\mathbb{R}}, \\ \chi \otimes a &\longmapsto a \cdot \omega_p \circ \chi \end{aligned}$$

we may view $V_{\mathbb{R}}$ as the root space of G with respect to T. In particular we may consider the highest weight ξ as well as half the sum of the positive roots η as elements of $V_{\mathbb{R}}$. Let $\leq$ denote the usual partial order on $V_{\mathbb{R}}$. Finally let W be, as before, the Weyl group of T and let z^{dom}, for any point $z \in V_{\mathbb{R}}$, be the unique dominant point in the W-orbit of z. We put

$${}^L T^\circ_{\xi,\mathrm{norm}} := \{\zeta \in {}^L T^\circ : \mathrm{val}(\zeta)^{\mathrm{dom}} \leq \eta + \xi\}.$$

Theorem 5.3. i. *${}^L T^\circ_{\xi,\mathrm{norm}}$ is an open K-affinoid subdomain of the dual torus ${}^L T^\circ$ which is preserved by the action of W.*

ii. *The Banach algebra $\mathcal{B}(G, \sigma_U)$ is naturally isomorphic to the ring of analytic functions on the quotient affinoid $W\backslash {}^L T^\circ_{\xi,\mathrm{norm}}$.*

Proof. [27], Prop. 2.4, Lemma 2.7, and the discussion before the remark in §6. □

For any given highest weight ξ the parameter space for our family of Banach space representations $B_{\xi,\zeta}$ therefore is $W\backslash {}^L T^\circ_{\xi,\mathrm{norm}}$. We emphasize that the pair

(ξ, ζ) should be viewed as consisting of a K-rational cocharacter $\xi \in X_*({}^L G^\circ)$ and a semisimple conjugacy class ζ in ${}^L G^\circ$.

In a second step we have to recognize this parameter space on the Galois side. This has its origin in a fundamental theorem about p-adic Galois representations ([6]) which asserts the existence of an equivalence of categories

$$\begin{array}{lll} \textit{Fon}: & K\text{-linear crystalline representations of } \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) & \xrightarrow{\sim} \quad \text{weakly admissible filtered } K\text{- isocrystals.} \end{array}$$

A filtered K-isocrystal is a triple $(D, \varphi, \mathrm{Fil}^\cdot D)$ consisting of a finite dimensional K-vector space D, a K-linear automorphism φ of D – the "Frobenius" –, and an exhaustive and separated decreasing filtration $\mathrm{Fil}^\cdot D = (\mathrm{Fil}^i D)_{i\in\mathbb{Z}}$ on D by K-subspaces. Note that the pair (D, φ) can be viewed as a Weil–Deligne group representation $\mathcal{W}_{\mathbb{Q}_p} \to \mathcal{W}_{\mathbb{Q}_p}/\mathcal{I}_{\mathbb{Q}_p} \to \mathrm{GL}(D)$ sending ϕ to φ. Let FIC_K denote the additive tensor category of filtered K-isocrystals. Weak admissibility is a certain condition on the relation between the filtration $\mathrm{Fil}^\cdot D$ and the eigenvalues of the Frobenius φ which we will not recall here.

Let $\mathrm{REP}_K({}^L G^\circ)$ denote the Tannakian category of all K-rational representations of ${}^L G^\circ$. Consider now any pair (ν, b) consisting of a K-rational cocharacter $\nu \in X_*({}^L G^\circ)$ and an element $b \in {}^L G^\circ$. We then have the tensor functor

$$\begin{aligned} I_{(\nu,b)} : \mathrm{REP}_K({}^L G^\circ) &\longrightarrow \mathrm{FIC}_K \\ (\tau, D) &\longmapsto (D, \tau(b), \mathrm{Fil}^\cdot_{\tau\circ\nu} D) \end{aligned}$$

with the filtration

$$\mathrm{Fil}^i_{\tau\circ\nu} D := \bigoplus_{j\geq i} D_j$$

defined by the weight spaces D_j of the cocharacter $\tau \circ \nu$. Borrowing a terminology from [20], Chap. 1, we make the following definition.

Definition 5.4. The pair (ν, b) is called weakly admissible if the filtered K-isocrystal $I_{(\nu,b)}(\tau, D)$, for any (τ, D) in $\mathrm{REP}_K({}^L G^\circ)$, is weakly admissible.

Suppose that (ν, b) is weakly admissible. Then we may compose $I_{(\nu,b)}$ with the inverse of the functor *Fon* and obtain a faithful tensor functor

$$\Gamma_{(\nu,b)} : \mathrm{REP}_K({}^L G^\circ) \longrightarrow \mathrm{Rep}_K^{\mathrm{con}}(\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p))$$

into the Tannakian category of all finite dimensional K-linear continuous representations of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$. By the general formalism of neutral Tannakian categories ([7]) the functor $\Gamma_{(\nu,b)}$ gives rise to a continuous homomorphism of groups

$$\gamma_{\nu,b} := \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \longrightarrow {}^L G^\circ(\overline{K})$$

which is unique up to conjugation in ${}^L G^\circ(\overline{K})$. Hence any weakly admissible pair (ν, b) determines an isomorphism class of "Galois parameters" $\gamma_{\nu,b}$. The connection to our parameter space from the first step is provided by [27], Prop. 6.1, as follows.

Theorem 5.5. *Suppose that $\eta \in X^*(T)$, let $\xi \in X^*(T)$ be dominant, and let $\zeta \in {}^L T^\circ(K)$; then there exists a weakly admissible pair (ν, b) such that ν lies in the ${}^L G^\circ(K)$-orbit of $\xi\eta$ and b has semisimple part ζ if and only if $\zeta \in {}^L T^\circ_{\xi,\mathrm{norm}}(K)$.*

We see that, given a pair (ξ, ζ) with $\xi \in X^*(T)$ dominant and $\zeta \in {}^L T^\circ_{\xi,\mathrm{norm}}(K)$ and assuming that $\eta \in X^*(T)$, we have on the one hand the conjecturally nonzero unitary Banach space representation $B_{\xi,\zeta}$ of G. On the other hand we have the Galois parameters $\gamma_{\nu,b}$ into ${}^L G^\circ(\overline{K})$ for all weakly admissible (ν, b) such that ν is conjugate to $\xi\eta$ and b has semisimple part ζ. This is the basis for our belief that these Galois parameters $\gamma_{\nu,b}$ essentially classify the topologically irreducible "quotient" representations of $B_{\xi,\zeta}$ (the quotation marks indicate that we want to allow for the possibility that the quotient map only has dense image). This would constitute an unramified p-adic Langlands functoriality principle. The technical assumption that $\eta \in X^*(T)$ is satisfied if G is semisimple and simply connected. But it is interesting to realize that it can be altogether avoided by working with a modification of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$. By local class field theory the group $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ has a (up to isomorphism) unique nontrivial central extension of the form

$$1 \longrightarrow \{\pm 1\} \longrightarrow \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)^{(2)} \longrightarrow \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \longrightarrow 1.$$

We now impose on our coefficient field K the slightly stronger condition that $\mathbb{Q}_p^\times \subseteq (K^\times)^2$. If $\varepsilon\colon \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \to \mathbb{Z}_p^\times$ denotes the cyclotomic character then we have a cartesian square

$$\begin{array}{ccc} \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)^{(2)} & \longrightarrow & \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \\ \downarrow{\scriptstyle \varepsilon_2} & & \downarrow{\scriptstyle \varepsilon} \\ K^\times & \xrightarrow{(\cdot)^2} & K^\times. \end{array}$$

The important point is that on $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)^{(2)}$ the cyclotomic character ε has the square root ε_2. If $\mathbb{D}$ denotes the protorus with character group $\mathbb{Q}$ then η always can be viewed as a K-rational cocharacter $\eta\colon \mathbb{D} \to {}^L T^\circ$ such that $\eta^2 \in X_*({}^L T^\circ)$.

The notion of weak admissibility extends to filtered K-isocrystals where the filtration is indexed by $\frac{1}{2}\mathbb{Z}$. As a consequence we may define weak admissibility for any pair (ν, b) where $\nu\colon \mathbb{D} \to {}^L G^\circ$ is a K-rational cocharacter such that $\nu^2 \in X_*({}^L G^\circ)$ and $b \in {}^L G^\circ$. Theorem 5.5 without the restriction on η remains true in this more general setting ([27], end of §6). Moreover, it is shown in [4] that the above construction of Galois parameters extends in the sense that any weakly admissible pair (ν, b) (of this more general kind) such that ν is conjugate to $\xi\eta$ gives rise to an isomorphism class of "Galois parameters"

$$\gamma_{\nu,b}\colon \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)^{(2)} \longrightarrow {}^L G^\circ(\overline{K}).$$

The picture can be made somewhat more precise in the case of the group $\mathrm{GL}_{d+1}(\mathbb{Q}_p)$. But first we remark that the reason for our assumption that $p^{1/2} \in K$

is to make the affinoid ${}^L T^\circ_{\xi,\mathrm{norm}}$ functorial on the category $\mathrm{REP}_K({}^L G^\circ)$. Even for an arbitrary group G the Satake isomorphism can be renormalized in such a way that it is defined over any finite extension $K/\mathbb{Q}_p$. The Banach algebra $\mathcal{B}(G, \sigma_U)$ then becomes isomorphic to the ring of analytic functions on the quotient by W of the affinoid

$$ {}^L T^\circ_\xi := \{\zeta \in {}^L T^\circ : (\mathrm{val}(\zeta) + \eta)^{\mathrm{dom}} \leq \eta + \xi\}. $$

Of course, the W-action now is a twisted version of the natural action. If the derived group of G is simply connected then one can go one step further and make, in addition, the point η integral. In the following we describe this in the case of the general linear group.

For the rest of this section $K/\mathbb{Q}_p$ is an arbitrary finite extension, and we let $G = \mathrm{GL}_{d+1}(\mathbb{Q}_p)$. We also let $U := \mathrm{GL}_{d+1}(\mathbb{Z}_p)$ and T be the torus of diagonal matrices. Our preferred choice of positive roots corresponds to the Borel subgroup of lower triangular matrices. For any $1 \leq i \leq d+1$ we let $\lambda_i \in T/U \cap T$ be the coset of the diagonal matrix having p at the place i and 1 elsewhere. We make the identification

$$ \begin{aligned} V_{\mathbb{R}} = \mathrm{Hom}(T/U \cap T, \mathbb{R}) &\longrightarrow \mathbb{R}^{d+1} \\ z &\longmapsto (z_1, \ldots, z_{d+1}) \text{ with } z_i := z(\lambda_i). \end{aligned} $$

The dominant weight $\xi \in X^*(T)$ is given by

$$ \begin{pmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_{d+1} \end{pmatrix} \longmapsto \prod_{i=1}^{d+1} g_i^{a_i} $$

for an increasing sequence $a_1 \leq \cdots \leq a_{d+1}$ of integers. In fact, $(a_1, \ldots, a_{d+1})$ is the point in $\mathbb{R}^{d+1}$ which corresponds to ξ under the above identification. Our other point η corresponds to

$$ \frac{1}{2}(-d, -(d-2), \ldots, d-2, d). $$

We now note that the point

$$ \tilde{\eta} := (0, \ldots, d) = \eta + \frac{1}{2}(d, \ldots, d) $$

is integral with a correcting summand which is fixed by W. Hence we may rewrite the definition of ${}^L T^\circ_\xi$ as

$$ {}^L T^\circ_\xi = \{\zeta \in {}^L T^\circ : (\mathrm{val}(\zeta) + \tilde{\eta})^{\mathrm{dom}} \leq \tilde{\eta} + \xi\}. $$

Finally, for ${}^L T^\circ(K) = \mathrm{Hom}(T/U \cap T, K^\times)$ we use the coordinates

$$ \begin{aligned} {}^L T^\circ &\longrightarrow (K^\times)^{d+1} \\ \zeta &\longmapsto (\zeta_1, \ldots, \zeta_{d+1}) \text{ with } \zeta_i := p^{i-1}\zeta(\lambda_i). \end{aligned} $$

With these identifications our map val corresponds to the map

$$\begin{aligned}(K^\times)^{d+1} &\longrightarrow \mathbb{R}^{d+1}\\ (\zeta_1,\dots,\zeta_{d+1}) &\longmapsto (\omega_p(\zeta_1),\dots,\omega_p(\zeta_{d+1})) - (0,\dots,d)\end{aligned}$$

and ${}^L T^\circ_\xi$ corresponds to the subdomain

$$\{(\zeta_1,\dots,\zeta_{d+1}) \in (K^\times)^{d+1} : (\omega_p(\zeta_1),\dots,\omega_p(\zeta_{d+1}))^{\mathrm{dom}} \le (a_1, a_2+1,\dots,a_{d+1}+d)\}$$

where now $(\,\cdot\,)^{\mathrm{dom}}$ simply means rearrangement in increasing order. Theorem 5.5 in this case amounts to the following.

Proposition 5.6. *For any $(\zeta_1,\dots,\zeta_{d+1}) \in (K^\times)^{d+1}$ the following are equivalent:*

i. *There is a weakly admissible filtered K-isocrystal of the form*

$$(K^{d+1}, \varphi, \mathrm{Fil}^\cdot K^{d+1})$$

such that $\zeta_1,\dots\zeta_{d+1}$ are the eigenvalues of φ and $(a_1, a_2+1,\dots,a_{d+1}+d)$ are the break points of the filtration $\mathrm{Fil}^\cdot K^{d+1}$.

ii. $(\omega_p(\zeta_1),\dots,\omega_p(\zeta_{d+1}))^{\mathrm{dom}} \le (a_1, a_2+1,\dots,a_{d+1}+d)$.

For any K-linear crystalline representation ρ of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ we call the break points of the filtration on $Fon(\rho)$ the Hodge–Tate coweights of ρ. Moreover, we say that ρ is K-split if all eigenvalues of the Frobenius on $Fon(\rho)$ are contained in K. Using the Colmez–Fontaine equivalence of categories we deduce from Proposition 5.6 the existence of a natural map

$$\begin{array}{l}\text{set of isomorphism classes of } (d+1)\text{-dimensional}\\ K\text{-split crystalline representations of } \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)\\ \text{all of whose Hodge–Tate coweights have multiplicity one}\end{array} \longrightarrow \dot{\bigcup_\sigma} \begin{array}{l}\text{set of } K\text{-valued}\\ \text{characters of}\\ \mathcal{B}(G,\sigma_U).\end{array}$$

In the limit with respect to K this map is surjective. Our earlier speculation means that the fiber in a point (ξ,ζ) in the right hand side should essentially parametrize the topologically irreducible "quotients" of $B_{\xi,\zeta}$.

For the group $G = \mathrm{GL}_2(\mathbb{Q}_p)$ the above picture was the original and basic insight of Breuil. The drastic simplification which occurs in this case is that the fibers of the above map have at most two elements and, in fact, only one element most of the time (whereas these fibers are infinite in general). Later Breuil and Berger were able in [2] to actually prove that the Banach space representations $B_{\xi,\zeta}$ in the case where the corresponding two dimensional crystalline Galois representation is irreducible indeed are nonzero, topologically irreducible, and admissible.

We finish by remarking that the content of this section can be developed for any base field L finite over $\mathbb{Q}_p$. We refer to [22] and [4] for the details.

References

[1] Amice, Y., Duals. In *Proceedings of the Conference on p-adic Analysis* (Nijmegen, 1978), Report 7806, Katholieke Universiteit, Nijmegen 1978, 1–15.

[2] Berger, L., Breuil C., Représentations cristallines irréductibles de $GL_2(\mathbb{Q}_p)$. Preprint, 2005.

[3] Borel, A., Automorphic L-Functions. In *Automorphic Forms, Representations, and L-Functions* (ed. by A. Borel, W. Casselmann), Part 2, *Proc. Symp. Pure Math.* **33** (2) (1979), 27–61.

[4] Breuil, C., Schneider, P., First steps towards p-adic Langlands functoriality. Preprint, 2006.

[5] Colmez, P., Série principale unitaire pour $GL_2(\mathbb{Q}_p)$ et représentations triangulines de dimension 2. Preprint, 2005.

[6] Colmez, P., Fontaine, J.-M., Construction des représentations semistable. *Invent. Math.* **140** (2000), 1–43.

[7] Deligne, P., Milne, J. S., Tannakian categories. In *Hodge Cycles, Motives, and Shimura Varieties* (ed. by P. Deligne, J. S. Milne, A. Ogus, K.-Y. Shih), Lecture Notes in Math. 900, Springer-Verlag, Berlin, New York 1982, 101–228.

[8] Diarra, B., Sur quelques représentations p-adiques de $\mathbb{Z}_p$. *Nederl. Akad. Wetensch. Indag. Math.* **41** (1979), 481–493.

[9] Dixon, J. D., du Sautoy, M. P. F., Mann, A., Segal, D., *Analytic Pro-p-Groups*. Cambridge Stud. Adv. Math. 61, Cambridge University Press, Cambridge 1999.

[10] Emerton, M., Jacquet modules of locally analytic representations of p-adic reductive groups I: Definitions and first properties. *Ann. Sci. École Norm. Sup.*, to appear.

[11] Féaux de Lacroix, C. T., p-adische Distributionen. Diplomarbeit, Köln 1992.

[12] Féaux de Lacroix, C. T., Einige Resultate über die topologischen Darstellungen p-adischer Liegruppen auf unendlich dimensionalen Vektorräumen über einem p-adischen Körper. Thesis, Köln 1997; *Schriftenreihe Math. Inst. Univ. Münster* (3) **23** (1999), 1–111.

[13] Fontaine, J.-M., Représentations ℓ-adiques potentiellement semi-stables. In Périodes p-adiques. *Astérisque* **223** (1994), 321–347.

[14] Frommer, H., The locally analytic principal series of split reductive groups. Preprintreihe SFB 478 available at http://wwwmath1.uni-muenster.de/sfb/about/publ/. Münster 2003.

[15] Harris, M., Taylor, R., *The geometry and cohomology of some simple Shimura varieties*. Ann. of Math. Stud. 151, Princeton University Press, Princeton, NJ, 2001.

[16] Henniart, G., Une preuve simple des conjectures de Langlands pour GL(n) sur un corps p-adique. *Invent. Math.* **139** (2000), 439–455.

[17] Kohlhaase, J., Invariant distributions on p-adic analytic groups. Thesis, Münster 2005.

[18] Lazard, M., Groupes analytiques p-adique. *Inst. Hautes Études Sci. Publ. Math.* **26** (1965), 389–603.

[19] Morita, Y., Analytic representatios of SL_2 over a $\mathfrak{p}$-adic number field, II. In *Automorphic Forms of Several Variables* (ed. by I. Satake, Y. Morita), Progr. Math. 46, Birkhäuser Boston, Inc., Boston, MA, 1984, 282–297.

[20] Rapoport, M., Zink, T., *Period Spaces for p-divisible Groups*. Ann. of Math. Stud. 141, Princeton University Press, Princeton, NJ, 1996.

[21] Schneider, P., *Nonarchimedean Functional Analysis*. Springer Monogr. Math., Springer-Verlag, Berlin 2002.

[22] Schneider, P., Teitelbaum, J., $U(\mathfrak{g})$-finite locally analytic representations. *Represent. Theory* **5** (2001), 111–128.

[23] Schneider, P., Teitelbaum, J., Banach space representations and Iwasawa theory. *Israel J. Math.* **127** (2002), 359–380.

[24] Schneider, P., Teitelbaum, J., Locally analytic distributions and p-adic representation theory, with applications to GL_2. *J. Amer. Math. Soc.* **15** (2002), 443–468.

[25] Schneider, P., Teitelbaum, J., Algebras of p-adic distributions and admissible representations. *Invent. Math.* **153** (2003), 145–196.

[26] Schneider, P., Teitelbaum, J., Duality for admissible locally analytic representations. *Represent. Theory* **9** (2005), 297–326.

[27] Schneider, P., Teitelbaum, J., Banach-Hecke algebras and p-adic Galois representations. Preprint, 2005.

[28] Tate, J., Number theoretic background. In *Automorphic Forms, Representations, and L-Functions* (ed. by A. Borel, W. Casselmann), Part 2, *Proc. Symp. Pure Math.* **33** (2) (1979), 3–26.

[29] Vigneras, M.-F., Modular representations of p-adic groups and of affine Hecke algebras. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 667–677.

Westfälische Wilhelms-Universität Münster, Mathematisches Institut, Einsteinstraße 62, 48149 Münster, Germany

E-mail: pschnei@math.uni-muenster.de

The algebraization of Kazhdan's property (T)

Yehuda Shalom*

Abstract. We present the surge of activity since 2005, around what we call the algebraic (as contrasted with the geometric) approach to Kazhdan's property (T). The discussion includes also an announcement of a recent result (March 2006) regarding property (T) for linear groups over arbitrary finitely generated rings.

Keywords. Property (T), spectral gap, cohomology of unitary representations, bounded generation, stable range, expanders, sum–product phenomena, finite simple groups.

1. Introduction

I. The objectives and setting. A discrete group is said to have *property* (T) if every isometric action of it on a Hilbert space has a global fixed point. This property (in an equivalent characterization), was introduced by Kazhdan in 1967 [62], as a means to establish its two consequences: being *finitely generated* and having *finite abelianization*, for lattices in "higher rank" simple algebraic groups. While originally property (T) appeared unexpectedly, during the 70s–80s it found various surprising applications, e.g., to the first explicit construction of expander graphs (Margulis), the solution to the so-called Banach–Ruziewicz problem (Rosenblatt, Margulis, Sullivan), and in operator algebras, to the first constructions of type II_1 factors with a countable fundamental group (Connes). Since the 90s, and particularly during this decade, the study of property (T) has seen further rapid developments, both in theory and in applications, and its perception has substantially been transformed. It is now a fundamental notion and a powerful tool in diverse areas of mathematics, ranging from representation theory (where it was born), ergodic theory and geometric group theory, to operator algebras and descriptive set theory.

An excellent account of the 70s–80s theory can be found in de la Harpe–Valette's influential book [50]; the developments of the 90s are presented in Valette's Bourbaki [104], and a comprehensive up-to-date exposition of the subject can be found in the outstanding forthcoming book by Bekka, de la Harpe and Valette [11]. Consequently, our purpose is not to present another general exposition. Rather, we discuss two different, rather opposite trends in the study of property (T), *geometric* and *algebraic* which, we believe, can be detected quite clearly in retrospect. Following a brief historical account of the former, we focus our attention here mainly on recent exciting developments of the latter, and announce the following new result:

*Supported by the Israeli Science Foundation.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Theorem 1.1. *Let R be any finitely generated commutative ring with* 1. *Denote by* $\mathrm{EL}_n(R) < \mathrm{GL}_n(R)$ *the group generated by the elementary matrices over R. Then for all* $n \geq 2 + \mathrm{Krull\,dim} R$, *the group* $\mathrm{EL}_n(R)$ *has Kazhdan's property* (T).

In particular, it follows from a non-trivial "$\mathrm{EL}_n = \mathrm{SL}_n$" result of Suslin [101], that for any $m \geq 0$, the group $\mathrm{SL}_n(\mathbb{Z}[x_1, \ldots, x_m])$ has property (T) when $n \geq m+3$. These groups are the first known *linear* Kazhdan groups outside the family of lattices. In fact a better result holds, covering non commutative rings as well, in which Bass' ring theoretic notion of the *stable range* of R replaces the Krull dimension. The proof of the theorem also reduces to a new treatment of the "classical" case, $R = \mathbb{Z}$. The reader interested primarily in Theorem 1.1 may wish to skip to Section 4 on first reading, where a sketch of its proof is presented.

II. The geometrization of property (T). As is well known by now, a lattice Γ in a simple algebraic group G over a local field k has property (T), unless k-rank $G = 1$, but including $G = F_4^{-20}$, $\mathrm{Sp}(n, 1)$ when $k = \mathbb{R}$ – cf. [11], [50]. Due to the celebrated results of Margulis (rank > 1) and Corlette and Gromov–Schoen (rank 1), these simple algebraic groups turn out to be also the ones whose lattices enjoy *super-rigidity*, hence a fortiori, these lattices are *arithmetic groups* (with $\Gamma = \mathrm{SL}_{n\geq 3}(\mathbb{Z})$ serving as outstanding examples). For a long time, the simultaneous appearance of property (T) and superrigidity–arithmeticity was only empirical, having very different origin. However during the 90s, beginning with the breakthrough of Corlette [31], it has become evident that the theory of harmonic maps should provide a unified explanation for the two phenomena, even though in practice this has been carried out primarily for archimedean k and co-compact Γ. Be that as it may, for two decades following Kazhdan's discovery, essentially no new constructions of Kazhdan groups were found, and the proofs of property (T) for the *arithmetic groups* were depending crucially on their being *lattices*. As the latter are intimately related with the special geometry of symmetric spaces or Bruhat–Tits buildings, as well as with arithmetic-algebraic objects, this highly rigid framework was naturally projected back to the general perception of property (T).

The first new constructions of Kazhdan groups were put forward in Gromov's seminal work [46], as quotients of co-compact lattices in the rank one Kazhdan Lie groups. Although this geometric method gave rise to a continuum of Kazhdan groups, it is above all an achievement of Gromov's hyperbolic group theory; from the point of view of property (T), the implicit constructions are "deformations" of existing ones. It is only in 1994 that the first explicit constructions of entirely new Kazhdan groups appeared, in a remarkable work of Cartwright, Młotkowski and Steger [26]. The groups constructed there act simply transitively on the vertices of certain "exotic $\tilde{A}_2$ buildings" introduced in [25], and for a natural generating subset, the *best* Kazhdan constant was computed. At the time, these groups seemed to form a "singular" class of "cousins" of standard lattices, and their original treatment was quite algebraic. However, it is now understood that they are outstanding representatives of the "geometrization of property (T)", an approach going back to Garland's seminal paper [44].

In 1973, Garland [44] established the first general results in what has later developed to become the "vanishing below the rank" principle. Loosely speaking, this asserts that for a simple algebraic group G over a local field k, the cohomology $H^i(G,\pi)$ vanishes for a wide class of representations π, as long as $i < k - \text{rank}G$. A similar statement is inherited by the discrete co-compact subgroups $\Gamma < G$. Although soon a complete algebraic theory had been established in this setting (cf. [13], [27], [64], [107]), it is the geometric approach of Garland that has lent itself to broad generalizations outside the linear framework, thereby giving birth, a decade ago, to the "geometrization of property (T)". At the heart of Garland's approach lies the idea that an appropriate bound on the norm of a *local* Laplacian, defined on the links of a complex on which Γ acts, leads to vanishing of cohomology. Since the above definition of property (T) is tautologically equivalent to the vanishing of $H^1(\Gamma,\pi)$ for any unitary Γ-representation π, this can be used to give an extremely useful, geometric criterion for the presence of property (T). An illuminating account of the remarkable path from the classical Hodge theory and Matsushima and Bochner type formulae in the theory of harmonic maps, through Garland's work, to Kazhdan's property (T), can be found in Pansu's [84]. As was pointed out to us by Lior Silberman, one can now present a particularly simple proof of the following resulting local criterion for property (T) (due to Ballmann–Świątkowski [8], Pansu [84], Żuk [111]), using the most basic Poincaré type inequality for Hilbert space valued functions on finite graphs:

Theorem 1.2. *Let X be a 2-dimensional simplicial complex on which the group Γ acts properly and co-finitely by automorphisms. Assume that each vertex and each edge of X is contained in some triangle. If for any vertex x, its link is a connected graph whose first positive eigenvalue is $> 1/2$, then Γ has property* (T).

See also [11, Ch. 5], [82], [108], and particularly [53, Theorem 6.4]. All of the aforementioned works, as well as [35], [36], furnish us with a rich and wild family of Kazhdan groups, well beyond the original distinguished class of arithmetic groups. Last, but not least in this direction, is Gromov's "random groups" paper [47], in which the geometrization approach to fixed point properties of groups culminates in the construction of remarkable groups, having property (T) among other important features; see also the related elaborations [45], [81], [100].

III. From geometrization to algebraization of property (T). The geometrization of property (T), when it applies, is a powerful tool. It is so sweeping that it typically yields a much stronger fixed point property, covering at least all isometric actions on non-positively curved manifolds (as in Theorem 1.2 above – cf. [53, Theorem 6.4]). Thus, it *intrinsically* cannot apply when dealing with such interesting groups as $\text{SL}_n(\mathbb{Z})$. Moreover, so far it has not produced a single example of a *linear* group which is not a standard arithmetic lattice. Related to this, the crude scissors of the geometric approach are currently helpless in dealing with delicate questions regarding expanding properties of *infinite families of finite groups*. It is exactly for these problems that

we shall see the advantages of the recent *algebraic approach* to property (T). Unlike the geometric one, which looks at the group "locally", and essentially as a purely geometric object, the algebraic approach relies heavily on precise global algebraic structure. It applies a finer spectral analysis, and generally offers a more individual, less collective treatment. Our main purpose in this exposition is to describe its recent developments and achievements.

Trying to "geometrize the algebraization" – a failure report. Before proving Theorem 1.1 above, an attempt was made, together with Donald Cartwright, Lior Silberman, and Tim Steger, to find a computer assisted proof of some cases treated by Theorem 1.1, using Theorem 1.2 applied to Cayley complexes associated with the group. More precisely, going over $\sim 10^6$ generating subsets, the computer tested Żuk's-"$\lambda_1 > 1/2$" criterion for property (T) in an improved version, taking generating subsets invariant under conjugation by a finite subgroup, and applying a corrected variation of Żuk's Theorem 8 in [112]. While as explained, the condition inherently cannot hold over rings like $\mathbb{Z}$ or $\mathbb{Z}[x]$, a priori there seems to be no reason why it should not be satisfied when working with $R = \mathbb{F}_p[x, y]$, let alone with $R = \mathbb{F}_p[x]$ (for which $\mathrm{SL}_3(R)$ is a lattice on a $\tilde{A}_2$ building, and has property (T)). The attempts (with $p = 2, 3$) failed. There indeed seems to be little intersection between the two approaches to property (T).

Acknowledgments. Thanks to J. Bourgain, P. de la Harpe, H. Helfgott, M. Kassabov, A. Lubotzky, A. Valette, and particularly to T. Steger, for helpful remarks and discussions on this exposition. Being completed essentially on our wedding anniversary March 30, the paper is dedicated to my wife Marina, with my love and appreciation.

2. First algebraization results: bounded generation and general rings

I. Kazhdan constants. One driving force in the study of property (T) is the determination of *explicit Kazhdan constants*. Recall (the well known Delorme–Guichardet theorem, cf. [50]) that a finitely generated group Γ has property (T) iff for every finite generating subset $S \subseteq \Gamma$ there exists some $\varepsilon > 0$, such that the following is satisfied: *If $\pi : \Gamma \to U(V)$ is any unitary Γ-representation on a Hilbert space V, for which there is some $v \in V$ with $\|\pi(g)v - v\| < \varepsilon\|v\|$ (such v is called (S, ε)-invariant), then there is some $0 \neq u \in V$ which is Γ-invariant.* Any $\varepsilon > 0$ is referred to as a *Kazhdan constant* for Γ, with respect to (and usually depending on) S.

Besides serving as a natural challenge, determining explicit Kazhdan constants makes quantitative many of the applications of property (T). It also makes available the next qualitative problem of *uniformity* of property (T) (more precisely, of the Kazhdan constant) over a family of groups, which reduces to the important theme of expander Cayley graphs for finite groups (discussed in Section 3 below). The particularly natural, intriguing case of $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$, with its "canonical" generating subset of $n^2 - n$

unit elementary matrices, was a problem raised by Serre in the 80s – see also the problem list in [50].

In [94] it was realized that Kazhdan's original proof of property (T) for higher rank algebraic groups, actually is (or can be made) effective. Thus, explicit and even optimal Kazhdan constants can be obtained for these groups (see also [10], [12], [80] in this direction). From there, by making quantitative Kazhdan's original argument that any lattice in a Kazhdan group is Kazhdan, explicit Kazhdan constants for each lattice can be obtained, based on some (soft) information on a fundamental domain of it. The latter being handled by standard reduction theory, this settles the issue *in principle*. In practice, however, matters are not quite as simple, and many basic questions regarding the asymptotic behavior of the Kazhdan constants are difficult to understand from this viewpoint. Although this is the solution one can hope for when dealing with the *family* of all lattices, for many individual arithmetic groups, such as $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$, it yields Kazhdan constants for generating sets of "geometric" rather than "algebraic" nature. From the more general perspective we are trying to pursue here, this is far from being the "right" solution, as it continues to treat the arithmetic groups as *lattices*, rather than approaching them as independent groups. Such an approach was indeed accomplished in [96], providing a solution to Serre's question above for $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$. In that paper a systematic use of two tools was initiated: the group theoretic notion of *bounded generation*, and the passage from standard rings like $\mathbb{Z}$, to *arbitrary* finitely generated rings, like $\mathbb{Z}[x]$, via a general relative property (T) result. The next two subsections introduce these two ingredients, and some of their earlier (yet still quite recent) combined applications. Further recent developments follow in subsequent sections.

II. Bounded generation

Definition 2.1. Let G be a group, and $\{H_i\}$ be a finite family of subgroups. We say that G is *boundedly generated* by $\{H_i\}$, if there exists some $M < \infty$, such that every $g \in G$ is a product of at most M elements, each belonging to some H_i. If the H_i are cyclic subgroups, we simply say that G is boundedly generated.

This notion, and the first non-trivial examples of it, came with the work of Carter–Keller [24], who showed that for the ring of integers $\mathcal{O}$ of any number field k, $\mathrm{SL}_{n\geq 3}(\mathcal{O})$ is boundedly generated. More precisely, they showed that this group is boundedly generated by the family of its $(n^2 - n)$ *elementary subgroups*, a property which makes sense when working over *any* ring R, and is then termed *bounded elementary generation*. Carter–Keller's result uses Dirichlet's theorem on primes in arithmetic progressions, and gives an explicit bound on M in terms of n and the discriminant ($2n^2 + 50$ works for $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$ – see also the friendlier account in [1], still not avoiding Dirichlet's theorem, where matters stand as of today). Bounded generation has since been studied further, especially for arithmetic groups and in relation to the congruence subgroup property (cf. [85], [109] and the references therein). Although it certainly deserves more attention, we shall not be able to discuss it out-

side the framework of property (T), except to mention the fundamentally important problem of deciding whether it is shared by any single *co-compact* lattice in a higher rank simple Lie group.

The relevance of bounded generation to questions around property (T) was first demonstrated by Colin de Verdière (cf. [30, Theorem 3.9]), and independently in [96]. The idea can be easily explained through the following simple, yet remarkably useful observation:

Lemma 2.2 (Bounded Generation Lemma). *Assume that G is boundedly generated by $\{H_i\}$. If an isometric G-action on a Hilbert space admits a fixed point for each H_i separately, then it admits a global fixed point.*

Proof. The existence of a fixed point for one H_i, implies that all its orbits are norm bounded (as the action is isometric). Therefore the G-action has bounded orbits, and the unique circumcenter of one such orbit is fixed by all of G. □

In reality, one often tries to argue more quantitatively, at the level of the *unitary representation*, in order to get an *explicit Kazhdan constant* for the group G. The general scheme goes as follows:

(i) Show that for any unitary G-representation with almost invariant vectors, there is an H_i-invariant vector for each i. Quantitatively, any ε-invariant vector v (with respect to a fixed generating set), is ε' close to an H_i-invariant vector, for each i separately.

(ii) Deduce from (i) that *the same* v is $2\varepsilon'$-invariant under *all* of H_i.

(iii) By choosing $\varepsilon > 0$ small enough, make $2\varepsilon' < 1/M$, where M is the bounded generation constant. Hence all of G moves the unit vector v by less than unit distance, and the circumcenter of Gv is a *non-zero* vector, invariant under all of G.

As we shall see, this scheme (with small variations) turns out to be extremely useful even when all the groups involved are *finite*, enabling one to "lift" Kazhdan constants from "smaller" to "larger" groups, upon having a precise structural algebraic information. The main spectral analysis lies in part (i), and it is this *relative property* (T) (of G with respect to H_i), to which we now turn our attention.

III. The relative property (T) over general rings. The relative property (T) is an important variant which is implicit in Kazhdan's original paper, and was first introduced by Margulis (cf. [74], [75]). In analogy with Delorme–Guichardet's equivalent characterization of property (T), Julissaint [54] established the following (see also de Cornulier's extension of this notion in [32]):

Definition 2.3. Let Γ be a discrete group and $N < \Gamma$ a subgroup. We say that the pair (Γ, N) has the *relative property* (T), if either one of the following equivalent conditions is satisfied:

(i) Any isometric Γ-action on a Hilbert space admits a fixed point for N.

(ii) There exists a finite subset $S \subseteq \Gamma$ and $\varepsilon > 0$ ("Kazhdan constants"), so that any unitary Γ-representation containing a (S, ε)-invariant vector, admits a non-zero vector invariant under N.

The outstanding example, used by Margulis in his first explicit construction of expanders [74], is the semi-direct product $\Gamma = \mathrm{SL}_2(\mathbb{Z}) \ltimes \mathbb{Z}^2$, with $N = \mathbb{Z}^2$. In the course of computing explicit Kazhdan constants for the finite representations of $\mathrm{SL}_3(\mathbb{Z})$, Burger [21] found explicit Kazhdan constants for this pair. While his method used, by means of unitary induction, the co-compact embedding $\mathbb{Z}^2 < \mathbb{R}^2$, a variant avoiding it was found in [96]. This seemingly technical issue turned out to be of importance, as it triggered the passage to working with *general finitely generated commutative rings*, thereby releasing a part of the theory from the burden of an ambient locally compact group. Kassabov observed that the commutativity of the ring multiplication operation is not required in the proof, and consequently we have [58], [96]:

Theorem 2.4. *Let R be any finitely generated ring with* 1. *Then* $(\mathrm{EL}_2(R) \ltimes R^2, R^2)$ (*and* $(\mathrm{SL}_2(R) \ltimes R^2, R^2)$ *when R is commutative*)*, has the relative property* (T)*, with explicit Kazhdan constants available, depending only on the number of generators of R.*

The main tool in the proof of the result is the spectral theorem for representations of abelian groups. By taking the spectral measure on the Pontrjagin dual $\widehat{R^2}$, corresponding to almost $\mathrm{EL}_2(R)$-invariant vectors, one gets a sequence of almost $\mathrm{EL}_2(R)$-invariant measures with respect to the dual action on $\widehat{R^2}$. It is then shown that such a sequence of measures cannot exist when they have "most" of their support "close" to (but excluding) $0 \in \widehat{R^2}$. That, however, would have been the case if the vectors in consideration were taken to be also *almost R^2-invariant*. The whole proof can be made quantitative, leading to explicit Kazhdan constants for the relative property (T). Without getting more technical, we note (anticipating the sum–product results to be discussed in Section 3.III below), the tension used here between the two algebraic operations of the ring.

IV. Bounded generation + relative (T) approach: first applications. It was shown in [96] that when $R = \mathbb{Z}$ in Theorem 2.4, one can take $\varepsilon = 1/10$ in Definition 2.3 (ii), for the generating set S consisting of the unit elementary matrices of $\mathrm{SL}_2(\mathbb{Z})$ and the standard basis of $\mathbb{Z}^2$. It is easy to see that every elementary subgroup $H_i \cong \mathbb{Z}$ of $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$ can be placed in a copy of $\mathbb{Z}^2$, which is normalized by some embedding of $\mathrm{SL}_2(\mathbb{Z})$ in $\mathrm{SL}_n(\mathbb{Z})$. Thus, the relative property (T) for $(\mathrm{SL}_2(\mathbb{Z}) \ltimes \mathbb{Z}^2, \mathbb{Z}^2)$ gives the explicit $\varepsilon = 1/10$ in step (i) of the quantitative scheme described following Lemma 2.2, for the generating set of $n^2 - n$ unit elementary matrices. Taking the Carter–Keller bounded generation estimate, one completes the last two steps of the scheme to get explicit Kazhdan constants for $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$, which decrease quadratically in n. Later, Kassabov realized in [56] that with additional effort, one can execute this scheme "more efficiently" for large n, to obtain Theorem 2.5 below. It is quite remarkable that while the unitary dual of $\mathrm{SL}_n(\mathbb{Z})$ (or any non-virtually abelian group for that matter), is entirely out of reach, ultimately these ideas give rise to an *explicit* determination of the *precise asymptotic behavior* of the Kazhdan constants of these groups, over all n:

Theorem 2.5. *Let ε_n denote the largest Kazhdan constant of the group* $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$ *w.r.t. the set of $n^2 - n$ unit elementary matrices. Then* $10^{-3}n^{-\frac{1}{2}} < \varepsilon_n < 2n^{-\frac{1}{2}}$.

A more careful analysis reduces the ratio between the two bounds to 60 as $n \to \infty$.

The successful treatment of $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$, together with the generality of Theorem 2.4, suggests that one should aim higher, at property (T) for other groups of similar type, notably $\mathrm{SL}_{n\geq 3}(\mathbb{Z}[x_1, \ldots, x_m])$. These groups are called "universal lattices" in [96], as they naturally surject onto many standard arithmetic lattices over both zero and positive characteristic local fields. Property (T) for them would thus account in a uniform manner for Kazhdan's property of very different arithmetic groups. Notice that if we knew that for $n \geq 3$ they were *boundedly elementary generated*, then the same strategy as for $\mathrm{SL}_n(\mathbb{Z})$ would establish property (T) for them as well. It was conjectured in [96] that property (T) should indeed be present among these groups, even though the question of bounded elementary generation has been open since its introduction in the context of K-theory by van der Kallen, in 1982 [55] (where it was also shown that bounded elementary generation over the ring $\mathbb{C}[x]$ does not hold). As will be seen in Section 4 below, the proof of Theorem 1.1 circumvents this delicate, still unsettled issue, by using a different cohomological proof which requires a more modest bounded generation property (which shows up when n is larger than the dimension of the ring $+1$).

The following variation on the same general scheme enables one to get other interesting consequences, where bounded elementary generation is less elusive. It is implicit in the proof of [96, Corollary 4], and subsumes the previous discussion as well, when the ring is taken with its discrete topology.

Theorem 2.6. *Fix $n \geq 3$. Assume that for some finitely generated ring R with* 1*, the group* $\mathrm{EL}_n(R)$ *is embedded* densely *in a topological group G, and the group G is boundedly generated by the* closure *of the embeddings of the elementary subgroups of* $\mathrm{EL}_n(R)$*. Then every* continuous *unitary G-representations with almost invariant vectors for* $\mathrm{EL}_n(R)$*, has a non-zero vector invariant under* all *of G. Moreover, an explicit bound on the Kazhdan constant and size of the Kazhdan set is available, depending only on the number of generators of R as a ring, and on the bounded generation estimate.*

We conclude this section with two rather different applications of this result.

Theorem 2.7. *For $n \geq 3$, the infinite dimensional loop group $G = L(\mathrm{SL}_n(\mathbb{C}))$ of all continuous maps from the circle to* $\mathrm{SL}_n(\mathbb{C})$*, has property* (T)*. More precisely, there is a finite subset $S \subseteq G$ and $\varepsilon > 0$, so that every continuous unitary G-representation with an (S, ε)-invariant vector, admits a non-zero invariant vector.*

This result from [96] is the first construction of an infinite dimensional Lie group, and in fact of any *non locally compact* topological group, with property (T) (more recent ones can be found in [9], [33], [78]). The second application of Theorem 2.6 is the following deeper result of Kassabov and Nikolov [61], giving the first positive

result towards Theorem 1.1 above. The relevant ambient topological group G here, is the profinite completion.

Theorem 2.8. *Fix $n \geq 3$. Then for any finitely generated commutative ring R, the group* $\mathrm{EL}_n(R)$ *has property* (τ)*, namely, the family of all its* finite *representations not containing an invariant vector, does not contain any (S, ε)-invariant vector (for an explicit finite subset S and $\varepsilon > 0$). In particular, property (τ) holds for the groups* $\mathrm{SL}_{n\geq 3}(\mathbb{Z}[x_1, \ldots, x_m])$.

The bounded generation result needed for Theorem 2.8 controls the "failure of the congruence subgroup property" over higher dimensional rings. This is obtained by proving a result of independent interest, on the bounded generation of $K_2(R)$ by products of Steinberg symbols $\{a, b\}$.

3. The algebraization of property (T) for finite groups: expanders

The general theme of expanders, which has attracted much attention and interest in computer science, combinatorics and group theory, needs by now no introduction (cf. [69], [90] and references therein, for more information, particularly in the directions we shall follow). When dealing with Cayley graphs, the relation to property (T) is fundamental: *A family of finite groups, each equipped with a generating set of uniformly bounded size, is an* expander*, iff their Kazhdan constants are all uniformly bounded away from* 0. For our purposes, and for the benefit of the interested non-specialist, we may simply regard the latter as our definition of expanders in the framework of Cayley graphs. The main general group theoretic problems in this setting can be put into two related, yet independent directions:

1. Existence. Given a family of finite groups G_i, can one find for each i a generating set $S_i \subseteq G_i$, making $\mathrm{Cay}(G_i, S_i)$ an expander family?

2. Independence. Given a family G_i for which 1 is answered positively, is it an expander family with respect to *all* generating subsets? *Random* ones?

Implicit here is the natural question (due to Lubotzky–Weiss [73]), settled *positively* only in recent work of Alon–Lubotzky–Wigderson [3] (following [86]), of whether being an expander family *depends in general* on the choice of generating subsets. This makes any positive results towards 2 of substantial interest. For completeness, we mention that Alon–Roichman [4] showed that any family of finite groups can be made expander using generating subsets of logarithmic size, and that some families, e.g. consisting of abelian groups, *cannot* be made expanders. A completely positive answer to 2 is still unknown for any infinite family of groups; an empirical support towards it for the family $\{\mathrm{SL}_2(\mathbb{F}_p)\}$ was provided by Lafferty–Rockmore in [65], [66].

In this section we describe the remarkable progress made on these two problems *during the last year* (2005), which turns out to be very closely related to the previously

discussed algebraization approach to property (T). Before getting into more details, we discuss briefly some relevant background, which shows striking parallelism to the early slow developments in constructing new Kazhdan groups (as discussed in the introduction), and may help to put this exciting progress in perspective.

I. Some background. Margulis' 1973 first explicit construction of expander graphs gives rise to a general class of so-called *"mother group expanders"*: Take a finitely generated ("mother"-)group Γ generated by a finite (symmetric) subset $S \subseteq \Gamma$, an infinite sequence of finite index normal subgroups $N_i \lhd \Gamma$, and consider $\mathrm{Cay}(\Gamma/N_i, \bar{S})$ (where $\bar{S}$ is the canonical projection of S). This automatically yields expanders when Γ has property (T) (interestingly, Gromov's recent random constructions [47] mark a way back from expanders to property (T)). However different, even "better" constructions can be obtained, by using other mother groups such as free groups. The necessary spectral gap property is always a highly non-trivial matter, and for more than two decades after Margulis' construction, the only known approach besides property (T) relied on deep number theoretic tools, around Selberg ("$\lambda_1 \geq 3/16$")–Ramanujan type estimates. All the groups Γ which were known to become mother groups for expanders, were *arithmetic lattices* (with subgroups $N_i \lhd \Gamma$ taken to be *congruence*). As outstanding examples, one should keep in mind the two families $\mathrm{SL}_2(\mathbb{F}_p)$ and $\mathrm{SL}_{n\geq 3}(\mathbb{F}_p)$, *always taken with the projection of a generating set of the corresponding mother group:* $\mathrm{SL}_2(\mathbb{Z})$ (the number theoretic approach), or $\mathrm{SL}_{n\geq 3}(\mathbb{Z})$ (the property (T) approach).

As will become clear, it is not a coincidence that the prolonged lack of progress here was so reminiscent of the one described in the Introduction, concerning new constructions of Kazhdan groups. Besides the little flexibility available by choosing the generators for the quotients Γ/N_i as projections of a subset $S_0 \subseteq \Gamma$ generating a *finite index* subgroup $\Gamma_0 < \Gamma$, no single other construction was known until a decade ago. See Lubotzky's "frustrated account" of this state of affairs in [70]. There the "1–2–3" test case problem was suggested, of proving that $\mathrm{SL}_2(\mathbb{F}_p)$ are expanders with respect to the projection of the set S_0 of elementary 2×2 matrices with ± 3 off the diagonal (while the usual mother group generators yield expanders only for ± 1 or ± 2).

In [92] and [93] appeared the first new constructions of subsets S_0 of the mother group Γ, generating an *infinite index* subgroup, whose projections to Γ/N_i remain expanders. The general principle put forward there, was that one can retain the expanding property when $\langle S_0 \rangle = \Gamma_0 < \Gamma$ has *infinite index*, as long as Γ_0 is "*close enough*" to Γ, in terms of a comparison between the spectral gaps of $(\ell^2-)\Gamma/\Gamma_0$ and $(\ell^2_0-)\Gamma/N_i$. The spectral gap here can be measured by means of norms of convolution operators in $\mathbb{C}[\Gamma]$, or by comparing the Riemannian λ_0 vs λ_1 eigenvalues in geometric settings. For example, one can always remain with expanders, when using $S_0 \subseteq \Gamma$ generating a *co-amenable* subgroup $\Gamma_0 < \Gamma$ (e.g. Γ_0 is normal, and $\Gamma/\Gamma_0 \cong \mathbb{Z}$, cf. the concrete example in [93], computationally analyzed in [67]). A more interesting implementation of this principle can be obtained when Γ is a (necessarily) free group for

which Γ/N_i are *Ramanujan graphs* à la Lubotzky–Phillips–Sarnak–Margulis. Their optimal spectral gap property alone, implies that one can find (non-constructively, though) in *every non-trivial normal subgroup* $\Gamma_0 < \Gamma$, a finite subset S_0 whose projection to the finite quotients Γ/N_i (whatever groups they are), is an expander. The approach to these results is functional analytic, applying such tools as compactness in weak topologies, and the Krein–Milman theorem. It is based on an intimate connection with the Banach–Ruziewicz type problem on the profinite completion of Γ. Similarly to Gromov's constructions of Kazhdan groups as quotients of hyperbolic Kazhdan groups, this approach suffers from the fundamental drawback of introducing only a "deformation" of previously existing constructions. It is in itself incapable of providing expanders independently.

A second class of new, self contained constructions, came later in Gamburd's [40], where the same direction of taking $S_0 \subseteq \mathrm{SL}_2(\mathbb{Z})$ generating an infinite index "large" subgroup, is pursued. Here "largeness" is interpreted by the Hausdorff dimension (at least $5/6$) of the limit set on the boundary. The result still fell short of dealing with the "1–2–3" question of Lubotzky mentioned earlier. Gamburd's work relied on previous ideas of Sarnak and Xue [91] (see also [34]), which, as explained in Subsection III below, play a fundamental role in the recent far reaching work of Bourgain–Gamburd [15]. The latter, which concerns the independence problem 2, together with the work discussed in the next subsection regarding the existence problem 1, changed dramatically the poor progress made around the group theoretical aspects of expanders, by the end of the last century.

II. Making the finite simple groups into expanders. The first "new generation" Cayley graph expanders, i.e., ones not obtained via the previously discussed "mother group" approach, were established in [96]: *For any fixed $n \geq 3$ and $m > 0$, when R varies over all the commutative finite rings generated by at most m elements, the family* $\mathrm{SL}_n(R)$ *forms an expander family*. This immediately follows from Theorem 2.6, as it is a simple matter to give a uniform bounded elementary generation estimate for these rings, depending only on n. For example, all finite fields are generated as a ring by one element, hence $\{\mathrm{SL}_n(F)\}$, for any fixed $n \geq 3$, is an expander.

A major bridge was, however, yet to be crossed: obtaining a uniform Kazhdan constant for SL_n over a fixed finite field, but with *n growing*. At first glance, this seems quite unapproachable by the methods discussed earlier. However, using a fundamental clever variation, Kassabov showed [58] that one can extend the existing technique to encompass the latter as well, by taking appropriate *non-commutative rings R* in Theorem 2.6. More precisely, if F is a finite field, and if we let $R_d = \mathrm{Mat}_d(F)$ be the standard $d \times d$ matrix ring, then it can be shown that R_d is generated by 3 elements for all d, that $\mathrm{EL}_3(R_d)$ have a uniform bounded elementary generation property over all d, and that $\mathrm{EL}_3(R_d) = \mathrm{EL}_3(\mathrm{Mat}_d(F)) \cong \mathrm{SL}_{3d}(F)$. Hence, by Theorem 2.6 the latter form an expander family. Since it is easy to see that for all n and all finite fields F, $\mathrm{SL}_n(F)$ is uniformly boundedly generated by embeddings of $\mathrm{SL}_{3d}(F)$ for $d = [n/3]$, the quantitative version of the Bounded Generation Lemma 2.2 establishes

the following result of Kassabov [58]:

Theorem 3.1. *$\{\mathrm{SL}_n(F) \mid n \geq 3,\ F$ is a finite field$\}$ can be made an expander.*

We shall next formulate a considerably more general statement (whose proof uses this result), however we mark Theorem 3.1 as the first significant indication that one might hope to cover essentially all finite simple groups (at least those of Lie type). This was accomplished very recently as an accumulation of works by Kassabov, Lubotzky and Nikolov (cf. [60] and its references):

Theorem 3.2. *Excluding the Suzuki groups, the family of all finite* (*non-abelian*) *simple groups can be made an expander.*

The result was conjectured in [6] without the exception of the Suzuki groups. By the classification of finite simple groups, the proof amounts to dealing with the family of alternating groups A_n, and with the groups of Lie type (the finitely many sporadic groups are of course negligible in such asymptotic questions). An illuminating account of the work towards the proof of this theorem can be found in the joint announcement [60] of the three authors; we shall only present here some highlights, emphasizing the intimate relations with the main theme of this exposition.

Finite simple groups of Lie type. To complete first the family $\{\mathrm{SL}_n(\mathbb{F}_q)\}$ for all n and $q = p^k$, we are left, by Theorem 3.1 above, with the case $n = 2$. This is done by Lubotzky [71] in the following way: in [72] these groups are made Ramanujan with respect to sets $S_k^{(p)} \subseteq \mathrm{SL}_2(\mathbb{F}_{p^k})$ of (unbounded !) size $p + 1$. The Ramanujan spectral gap enters only is showing that they yield uniform Kazhdan constants for $\mathrm{SL}_2(\mathbb{F}_{p^k})$, over all k and p (yet with unbounded size of generating sets). However the specific construction in [72] is of use, in showing that for any p and k there is an element $g_k^{(p)} \in \mathrm{SL}_2(\mathbb{F}_{p^k})$ so that $S_k^{(p)} \subseteq \mathrm{SL}_2(\mathbb{F}_p) \cdot g_k^{(p)} \cdot \mathrm{SL}_2(\mathbb{F}_p)$. Since all the $\mathrm{SL}_2(\mathbb{F}_p)$'s can be made uniformly expanders with two generators, adding to those the $g_k^{(p)}$ and using the quantitative version of the Bounded Generation Lemma 2.2, yields expanding generating sets of three elements.

Once the case of $\{\mathrm{SL}_n(\mathbb{F}_q)\}$ has been settled, the quantitative version of the Bounded Generation Lemma 2.2 completes the treatment of all finite simple groups of Lie type, excluding the Suzuki groups, using the following two results:

Theorem 3.3. (1) (Nikolov [79]). *Every finite simple group G of* classical *type is a product of at most $M = 200$ conjugates of a subgroup H which is a* (*central*) *quotient of $\mathrm{SL}_n(\mathbb{F}_q)$, for some n and q.*

(2) (Lubotzky [71]). *Excluding the Suzuki family, for any family $X_r(q)$ of finite simple groups associated with a group X of* Lie type (*twisted or untwisted*) *and fixed rank r, there exists a constant M such that the statement in* (1) *holds* (*with $H \cong (P)\,\mathrm{SL}_2(\mathbb{F}_q)$*).

Since there are only finitely many families of finite simple groups not covered by (1), together with (2) Theorem 3.2 follows, excluding the alternating groups. The

difficulty in treating the Suzuki groups arises from the fact that the only simple groups they contain belong to that same family. The proof of (1) involves a detailed analysis of the subgroup structure of these groups. Although it seems that a more delicate treatment in this spirit should cover (2) as well, Lubotzky appeals instead to a model theoretic approach developed by Hrushovski and Pillay [52], which enables one to deduce the result by a kind of dimension argument, as if working over an algebraically closed field. However, we note that one can prove this result with less sophisticated model theoretic tools, by appealing to the *first order logic compactness theorem*. See [109] for a different useful relation between the latter theorem and bounded generation.

Symmetric groups. For the proof of Theorem 3.2 we are left with the family A_n (or equivalently S_n), which is the most intriguing and challenging among the groups covered in Theorem 3.2. One reason is that unlike the other families $X_r(q)$, it is easy to find (natural) bounded generating subsets for S_n, which make them *non-expanders*. Additionally, a simple cardinality computation, coupled with some basic knowledge of the subgroup structure of S_n, shows that a similar bounded generation strategy, using embeddings of $\mathrm{SL}_n(F)$, will not work here. Unfortunately, in this confined exposition we cannot do justice to the brilliant (and quite technical) work of Kassabov [57], who showed that S_n can be made an expander. Besides the original paper [57], see also the announcement [59] (an elaborate account), or the previously mentioned [60] (a less technical one). Below are only some highlights.

Kassabov proves his theorem by dividing the irreducible representations of S_n into two classes, according to whether the corresponding partition of n has the first row "small" (first class), or "large" (second class), and showing the uniform spectral gap in each class independently. This method is inspired by similar previous ideas of Roichman [89], whose work [88] is also applied in the analysis of the first class, to show that the Kazhdan constant of large (unbounded) sets $F_n \subseteq S_n$, consisting of "nearly all" elements in a suitable conjugacy class of S_n, is uniform. Although the sizes of these F_n are unbounded, Kassabov is able to confine them in a bounded product of uniformly expanding subgroups ($\cong$ products of $\mathrm{SL}_d(\mathbb{F}_2)$). Thus, using the quantitative version of the Bounded Generation Lemma 2.2, he obtains bounded Kazhdan sets for the representations in the first class. The argument for the second is entirely different: those representations are contained in $\ell^2(S_n/S_m)$, for appropriate m which is "close enough" to n, in order to imply strong transitivity (or fast mixing) for the action in that space. The precise argument yields uniform Kazhdan constants only for the family S_k with $k \sim 2^{18l}$, $l = 1, 2 \ldots$. The general case follows by showing that one can *uniformly boundedly generate* each S_n by embeddings of S_k for k of this type.

III. Uniform expansion over different generating sets. In a recent impressive achievement [15], Bourgain and Gamburd established the following:

Theorem 3.4. (1) *For any subset* $S \subseteq \mathrm{SL}_2(\mathbb{Z})$ *not generating a virtually cyclic subgroup,* $\mathrm{Cay}(\mathrm{SL}_2(\mathbb{F}_p), \overline{S})$ *is an expander family* (*for p large enough*).

(2) *Fix $k \geq 2$. As $p \to \infty$, an independent uniform random choice of k elements in each* $\mathrm{SL}_2(\mathbb{F}_p)$ *makes with probability* $\to 1$ *the* (*undirected*) *Cayley graphs into an expander.*

(3) *Fix any $c > 0$. If for every p a symmetric generating set $S_p \subseteq \mathrm{SL}_2(\mathbb{F}_p)$ is chosen, so that girth* $\mathrm{Cay}(\mathrm{SL}_2(\mathbb{F}_p), S_p) \geq c \cdot \log p$, *then these graphs form an expander.*

The heart of the result lies in part (3), which easily implies (1), and using the "random logarithmic girth" result established in [41], immediately implies (2). The proof of the theorem borrows key ideas from the approach introduced by Sarnak and Xue [91], incorporating two main ingredients: I. High multiplicity of the (bad) eigenvalues in the regular representation of $\mathrm{SL}_2(\mathbb{F}_p)$, stemming from Frobenius' classical result that the smallest dimension of a non-trivial representation of this group ($\frac{p-1}{2}$), is large relative to its size ($\sim p^3$) – i.e., there is a uniformly positive logarithmic ratio between the two (unlike the symmetric groups, for example). II. An upper bound on the number of returns to the identity for random walks of length up to logarithmic order of the group.

While previously, the upper bound in II was obtained by translating the problem into a Diophantine one, the generality of the Bourgain–Gamburd result is made possible by using instead tools from *additive combinatorics*. These include a non-commutative version of the Balog–Szemeredi–Gowers Lemma due to Tao [102], and notably Helfgott's recent breakthrough, discussed below, which capitalizes on *sum–product* results. Other aspects of Bourgain–Gamburd's work involve algebraic inputs, such as Frobenius' result above, and the precise subgroup structure of $\mathrm{SL}_2(\mathbb{F}_p)$. However, it is actually in the fascinating theme of sum–product phenomena, that one finds a rather striking similarity with previously discussed algebraization methods for property (T), and a conceptual explanation of how rich algebraic structure may lead to expansion properties. We shall try to shed some light on this fundamental ingredient, beginning with the recent pioneering result of Helfgott [51]:

Theorem 3.5. *Let $S \subseteq \mathrm{SL}_2(\mathbb{F}_p)$ be any generating set. Then* $\mathrm{Cay}(\mathrm{SL}_2(\mathbb{F}_p), S)$ *has diameter $\leq K(\log p)^c$, where the constants K, c are absolute.*

This result is weaker than "expansion" (in which case $c = 1$), and a similar statement (with $\log p$ replaced by $\log |G|$) is conjectured by Babai [5] to hold uniformly over all finite simple groups. However, it is here that for the first time, the barrier of handling uniformly independent generating subsets is crossed. The proof of Theorem 3.5 is a direct consequence of the following:

Key Proposition. *Let p be a prime and $A \subseteq \mathrm{SL}_2(\mathbb{F}_p)$. Then:*

(a) (Small sets) *If A is not contained in a proper subgroup, and $|A| < p^{3-\delta}$ with $\delta > 0$, then $|A \cdot A \cdot A| > c|A|^{1+\varepsilon}$, where $c, \varepsilon > 0$ depend only on δ.*

(b) (Large sets) *Assume A is not contained in any proper subgroup, and $|A| > p^{\delta}$, $\delta > 0$. Then there is an integer k depending only on δ, such that $(A \cup A^{-1})^k = \mathrm{SL}_2(\mathbb{F}_p)$.*

Theorem 3.5 follows immediately by first applying a constant number of set multiplications (depending only on c, ε), so that (a) starts giving exponential growth, and then applying (a) followed by (b). The latter is a major ingredient in Bourgain–Gamburd's work. Its proof when δ is close to 3 (e.g. $\delta > 8/3$) requires soft Fourier analysis, hence matters rest primarily on (a). We next discuss it, remarking first that the appearance of $|A \cdot A \cdot A|$ rather than $|A \cdot A|$ is necessary; consider e.g. $A = H \cup \{g_0\}$, where H is a subgroup.

Sum–product phenomena and expansion. Besides standard (by now) tools from additive combinatorics, such as the Balog–Szemeredi–Gowers theorem, and properties of Ruzsa distances, the proof of (a) in the Key Proposition makes crucial use of powerful *sum–product phenomena*. These arise in works of Bourgain, Glibichuk, Katz, Konyagin, Tao, and originally involved also subtle arithmetic techniques originated from (Stepanov's elementary proof of) Weil's work on the Riemann hypothesis over finite fields (the latter part is relevant only to Helfgott's and not to Bourgain–Gamburd's work, due to their logarithmic girth assumption). See [14], [18], [19], [63], Section 2.8 in [103], and the references therein, for further details, including the interesting intimate connections with work (notably by Bourgain), on the ring conjecture, Kakeya problem, and exponential sum estimates. For our purposes, it suffices to state the following (cf. [51]):

Theorem 3.6 (Sum–Product). *Fix $\delta > 0$. Then for any subset $A \subseteq \mathbb{F}_p - \{0\}$ with $C < |A| < p^{1-\delta}$, we have*

$$\max\{|A \cdot A|, |A + A|\} > |A|^{1+\varepsilon}$$

where $C, \varepsilon > 0$ depend only on δ.

An analogous result over the integers was first established by Erdös–Szemeredi [37]. Very recently, a simplified proof of the theorem was found by Tao (see Theorem 2.52, Corollary 2.55 in [103]). A similar statement holds for an arbitrary finite field F (taking into account the presence of subfields). Results accounting for a small growth of a set under applying internal arithmetic operations by an obvious algebraic structure (arithmetic progression, subring, subfield) capturing most of its mass, go back to Friemann's classical theorem (cf. [103, Ch. 2]). In fact, under the same heading one can also include the recent far reaching uniform exponential growth results of Eskin–Mozes–Oh [38], and Breuillard–Gelander [20], for infinite, finitely generated linear groups. In the latter, as in the proof of Theorem 3.6, one first shows that some algebraic operation on the set A yields a set with the desired growth property, and then deduces the result back for A itself. However, quite surprisingly, in the proof of Theorem 3.6 it is actually the *latter* step which is *trickier* (for instance, it can happen that $|A \cdot A + A \cdot A| \sim |A|^2$, but $|A \cdot A|, |A + A| < 2|A|$). Returning to Helfgott's Theorem 3.5 above, since matrix multiplication in $\mathrm{SL}_2(\mathbb{F}_p)$ encodes the addition and multiplication in $\mathbb{F}_p$ together, the relevance of sum–product results to Theorem 3.5 is not a surprise *a posteriori*. In practice, the proof applies Theorem 3.6

to the traces of the elements of a set, showing that the sizes of a set and its set of traces "keep track of one another".

If one wants to pin down the source of the expansion in both Bourgain–Gamburd and Helfgott results, it is the sum–product phenomenon which gives the best quick answer. It is quite interesting to examine in this light the proof of relative property (T) for general rings (Theorem 2.4 above), which is the departure point for most results in the algebraization of property (T). Its proof capitalizes on the same ring theoretic phenomenon, where sets (or measures) which are "almost invariant" under the combined ring operations, must be "degenerate" (e.g., must assign mass to 0).

The Archimedean spectral gap analogue. In this exposition we can only mention the second companion work by Bourgain–Gamburd [16], which also deals with establishing a uniform spectral gap, and involves some similar ingredients, this time working over $\mathbb{R}$ or $\mathbb{C}$. It is shown there that a certain "non-commutative Diophantine condition" on (the group generated by) a finite set $S \subseteq \mathrm{SU}(2)$, implies that its induced action on the zero mean functions $L^2_0(\mathrm{SU}(2))$ has a spectral gap. This Diophantine condition was introduced in previous (weaker) results in this direction by Gamburd–Jakobson–Sarnak [42], who also showed that it is automatically satisfied if all elements in S have algebraic traces. See these two papers, as well as [69], [90], for more on the history of the problem, which may be viewed as a quantitative version of the (positive solution to the) Banach–Ruziewicz problem, pertaining to a certain uniqueness property of the Lebesgue measure on the 2-sphere.

The bulk of this companion work by Bourgain–Gamburd consists in establishing a "statistical" analogue of Helfgott's Key Proposition above [16, Proposition 1]. Its proof replaces the sum–product Theorem 3.6 by an approach originating from (and improving on) Bourgain's work [14, Theorem 0.3], towards the ring problem. This work of Bourgain–Gamburd turns out to be more involved than the one on expanders, a fact which may seem surprising in view of past experience with these parallel problems. There are, however, two conceptual explanations why, when dealing with more "generic" (or less "special") finite subsets, one should expect more difficulties here. Firstly, while in the real topology "bad" sets can be continuously approached by "good" ones, such phenomenon cannot occur in the non-archimedean case. Even more importantly, the right analogue of the (ideal) SU(2) result, would be showing a uniform expansion for (topologically generating) finite subsets of the compact group $\mathrm{SL}_2(\mathbb{Z}_p)$ (p-adic integers), which, in return, is *equivalent* (see [92]) to showing that for any fixed k, *all* choices of k-generator subsets, independently in each $\mathrm{SL}_2(\mathbb{Z}/p^i\mathbb{Z})$, form expanders. Now, it may be expected that the latter property turns out more subtle than making $\mathrm{SL}_2(\mathbb{F}_p)$ with varying p into expanders. In the latter case, it is not a priori clear if and how the questions over different primes p relate. One may hope (perhaps naively though), that as it often happens, the "right" solution (spectral gap here) for one prime, would work uniformly over all primes. In contrast, in the case of $\mathrm{SL}_2(\mathbb{Z}/p^i\mathbb{Z})$ with i increasing, establishing a spectral gap bound for a given i automatically yields the same bound for smaller i's (as the canonical quotient map induces an inclusion at the L^2 level). Hence one faces more difficult tasks as i grows.

In fact, remarking on the preceding paragraph (in a previous draft of this paper), Jean Bourgain has informed us of substantial progress achieved recently towards a generalization of Theorem 3.4, covering uniformly essentially all $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$. He noted that the method when $q = p^i$ is indeed closer to the one in the SU(2) case, and may be viewed as its p-adic analogue – see below, and [17], where this generalization is involved. The higher rank cases $\mathrm{SL}_{n\geq 3}$ should naturally be addressed as well. Whether they turn out to be as demanding, remains to be seen.

Finally, we remark that Theorem 3.4 (with its extension mentioned above) has very recently been applied by Bourgain–Gamburd–Sarnak, to develop a combinatorial sieve method for primes and almost primes, on orbits of various subgroups of $\mathrm{GL}_n(\mathbb{Z})$ as they act on $\mathbb{Z}^n$. Unlike the more familiar case of sieving in $\mathbb{Z}^n$, in this setting the expander property plays a critical role. See the announcement [17] for further details.

4. Reduced cohomology and property (T) for elementary linear groups

I. Definition and basic properties of reduced cohomology. The study and use of reduced cohomology in relation to property (T) was first pursued in [95], motivated by establishing rigidity results for lattices in products of groups. We shall need it in Subsection III for the proof of Theorem 1.1, and in the next subsection some previous applications of it are mentioned. Recall first the basic correspondence between (affine) isometric actions of a group Γ on a Hilbert space V, and first cohomology. Any such action is of the form $\rho(\gamma)v = \pi(\gamma)v + b(\gamma)$, where π is a unitary Γ-representation on V, and the affine part $b\colon \Gamma \to V$ satisfies the 1-cocycle identity corresponding to π (the identity sufficient and necessary to make ρ an *action*). Fixing π, the set of all such b is a vector space, denoted $Z^1(\Gamma, \pi)$, and those "trivial" elements of the form $b_v(\gamma) = v - \pi(\gamma)v$, the coboundaries, form a subspace denoted $B^1(\Gamma, \pi)$. It is immediate that the Γ-action on V has a fixed point (v_0) iff the corresponding 1-cocycle b is a coboundary (b_{v_0}). We define the quotient space $H^1(\Gamma, \pi) = Z^1(\Gamma, \pi)/B^1(\Gamma, \pi)$, and can now consider also the topological version of it. Namely, fixing π, introducing the topology of pointwise convergence on the space $Z^1(\Gamma, \pi)$ makes it a Frechet space, in which B^1 is not always closed. Forming its closure, the first *reduced cohomology* $\overline{H}^1(\Gamma, \pi)$ can now be defined as $Z^1/\overline{B}^1$. The following analogous relation between fixed points of affine actions and coboundaries, can be verified easily:

Lemma 4.1. *Given $b \in Z^1(\Gamma, \pi)$, we have $b \in \overline{B}^1(\Gamma, \pi)$ iff the corresponding affine action $\rho(\gamma)v = \pi(\gamma)v + b(\gamma)$ has almost fixed points in the metric sense, namely, for every finite $S \subseteq \Gamma$, and $\varepsilon > 0$, there is $v \in V$ with $\|\rho(\gamma)v - v\| < \varepsilon$ for all $\gamma \in S$.*

All the notions and results here extend naturally to the class of second countable locally compact groups. The discussion above shows that the characterization of property (T) in terms of vanishing of usual cohomology is a tautology if one takes the fixed point property as a definition. However, the following characterization in

terms of the (generally smaller) reduced cohomology is less transparent, and holds only under the assumption of finite generation (or compact generation, in the locally compact setting):

Theorem 4.2 ([95]). *Let Γ be a finitely generated group without property* (T). *Then there exists some unitary Γ-representation π, with $\overline{H}^1(\Gamma, \pi) \neq 0$. Moreover, one can find such π which is irreducible.*

II. Some applications of the reduced cohomology

(1) Theorem 4.2 implies the existence of an irreducible π with $H^1(\Gamma, \pi) \neq 0$, as conjectured by Vershik–Karpushev [106]. Together with [106, Theorem 2] (see Louvet's [68] for a detailed exposition) it implies: *A discrete group Γ has property* (T) *iff it is finitely generated, has finite abelianization, and it does not admit any non-trivial irreducible unitary representation not separated* (*in the Fell topology*) *from the trivial representation*.

(2) As an application of the proof of Theorem 4.2, it is shown in [95] that every finitely generated Kazhdan group is a quotient of a *finitely presented* Kazhdan group (answering questions of Grigorchuk and of Żuk; the result is generalized by Fisher–Margulis [39] to locally compact groups). However, the existence seems entirely non-constructive, and there are concrete interesting groups which would be of interest to understand in this regards – see II in Section 5 below.

(3) Although in general, the existence of the *irreducible* cohomological π in Theorem 4.2 is non-constructive, somewhat surprisingly, in many cases one can actually classify all such π, and show that there are only *finitely many* of them. It may seem particularly unexpected that this finiteness phenomenon appears among *amenable* groups. For example, this is the case for all polycyclic (or lamplighter) groups, a result from [98] shown there to have applications in geometric group theory (e.g., any group quasi-isometric to a polycyclic group has a finite index subgroup with infinite abelianization). Martin [77] showed that all connected locally compact groups also have only finitely many such representations π.

(4) Inspired by Margulis' remarkable strategy in proving the normal subgroup theorem for higher rank lattices, as well as by more recent beautiful work of Burger–Mozes (cf. [22]), the following result was completed very recently: over sufficiently large finite fields, *any irreducible Kac–Moody group of non-affine* (*and non-spherical*) *type, has a finite index commutator subgroup, which is a* (*finitely generated*) *simple group, modulo its finite center*. Building on fundamental work of Rémy, the proof consists of three entirely independent results on any quotient of the simple group by a non-trivial normal subgroup: it is *Kazhdan* [95], *amenable* (Bader–Shalom [7]), and *infinite* (Caprace–Rémy [23]), classes of groups which do not intersect. The proof of the first relies crucially on the reduced cohomology of (infinite dimensional) unitary representations. Incidentally, we remark that while in virtually all of the applications of property (T), an appropriate Kazhdan group comes to the rescue, the ones made through normal subgroup theorems are of quite unique nature, as throughout the proof

no (non-trivial) Kazhdan group appears.

III. Sketch of proof of Theorem 1.1. Fix $n \geq 3$ (note that if $\dim R = 0$, by finite generation R must be finite). We proceed via the following steps:

1. Setting and notation. Set $\Gamma = \mathrm{EL}_n(R)$, and define the following subgroups: $\Lambda =$ matrices whose first row and first column begin with 1 and have 0s elsewhere, $N_1, N_2 \cong R^{n-1}$ the subgroups sitting in the upper row and left column with a "common" 1 at the upper left corner. Notice that Λ normalizes each one of the N_i's. The standard Steinberg commutator relations show that N_1 and N_2 together generate Γ. In fact, letting $r_1 = 1, r_2, \ldots, r_k$ be generators of R as a ring, the set S of all elementary matrices belonging to one of the two N_i's, having one of the r_j as the only non-zero element off the diagonal, forms a finite generating set for Γ.

Finally, for an isometric Γ-action ρ on a Hilbert space V_ρ and $v \in V_\rho$, denote

$$\delta_S(v) = \max\{\|\rho(s)v - v\| \mid s \in S\}, \quad \delta_S(\rho) = \inf\{\delta_S(v) \mid v \in V_\rho\}.$$

2. Reduced cohomology. Assume that Γ does not have property (T). We argue to get a contradiction. By Theorem 4.2 and Lemma 4.1 above, there exists some isometric Γ-action ρ on a Hilbert space V_ρ, with $\delta_S(\rho) > 0$. By rescaling we may assume that $\delta_S(\rho) \geq 1$. Denote by $\mathcal{A}$ the set of all isometric Γ-actions ρ with $\delta_S(\rho) \geq 1$.

3. Relative property (T) – the spectral ingredient. By (an obvious extension of) Theorem 2.4, for each i the pair $(\Lambda \ltimes N_i, N_i)$ has the relative property (T). Consequently, by the equivalence in Definition 2.3, the following infimum is not taken over the empty set:

$$d = \inf\{\|v^1 - v^2\| \mid v^i \in V_\rho \text{ with } \rho \in \mathcal{A}, \text{ and } \rho(N_i)v^i = v^i \text{ for } i = 1, 2\}.$$

4. Attaining d through a limiting process. Let $\rho_n \in \mathcal{A}$ and $v_n^i \in V_{\rho_n}^{N_i}$ with $\|v_n^1 - v_n^2\| = d_n \to d$. We may assume that $d_n < d + 1$, and this gives for all n $\delta_S(v_n^1) < 2(d+1)$, since v_n^1 is fixed by the N_1-generators of S, and a vector of distance at most $d + 1$ from it is fixed by the N_2-generators of S. This uniform bound implies (using an ultra-product argument as in [39], or a negative definite kernel argument as in [95]), that a subsequence of the actions (ρ_n, V_{ρ_n}), pointed at v_n^1, converges to an isometric Γ-action on a Hilbert space (ρ_∞, V_∞), with two points $v^i \in V_\infty^{N_i}$ satisfying $\|v^1 - v^2\| = d$. One shows that indeed $\rho_\infty \in \mathcal{A}$, hence d defined in step 3 is attained. Notice that $d \neq 0$, for otherwise v^1 is fixed by Γ, contradicting $\rho \in \mathcal{A}$.

5. Can assume π_∞ has no invariant vectors. Write $\rho_\infty(\gamma)v = \pi_\infty(\gamma)v + b_\infty(\gamma)$, where π_∞ is the (unitary) linear part (see the discussion at the beginning of the previous subsection). Decompose orthogonally $\pi_\infty = \pi_0 \oplus \pi_1$ where $\pi_0 = \pi^\Gamma$, and correspondingly, $b_\infty = b_0 + b_1$. Being a 1-cocycle for a trivial Γ-action, b_0 is an additive character, and since Γ is perfect (as it is generated by commutators), $b_0 = 0$. Replacing ρ_∞ by $\rho'_\infty(\gamma)v := \pi_1(\gamma)v + b_1(\gamma)$ yields the required reduction.

6. A fixed point for Λ via a geometric argument. Assume that for some $\lambda_0 \in \Lambda$, $w^1 := \rho_\infty(\lambda_0)v^1 \neq v^1$, and denote $w^2 = \rho_\infty(\lambda_0)v^2$. We will show this to be impossible. As Λ normalizes N_i, we have $v^i, w^i, u^i := \frac{1}{2}(v_i + w_i) \in V_\infty^{N_i}$, for $i = 1, 2$. Because λ_0 is an isometry, $\|v^1 - v^2\| = \|w^1 - w^2\| = d$, while by the definition of d as infimum, $\|u^1 - u^2\| \geq d$. By a standard convexity argument this is possible only if $v^1 - v^2 = w^1 - w^2$, and hence $v^1 - w^1 = v^2 - w^2 \neq 0$. But since v^1, w^1 are N_1-fixed, the left hand side is a non-zero *vector* invariant under the *linear* action of N_1, and similarly for $v^2 - w^2$ w.r.t. N_2. Since $\langle N_1, N_2 \rangle = \Gamma$, it follows that this common non-zero vector is Γ-invariant, contradicting the reduction made in step 5. Thus, $\rho(\Lambda)v^1 = v^1$.

7. Finishing with bounded generation and the stable range. A fundamental result of Bass (cf. Theorem 4.1.14 in [49]), asserts that if R is a commutative Noetherian (in particular, by Hilbert's basis theorem, if it is a finitely generated) ring, then for any $n \geq 2 + \text{Krull dim} R$, the ring R satisfies the following property: For every $a_1, \ldots, a_n \in R$ such that $a_1 R + \cdots + a_n R = R$ (such an n-tuple is called *unimodular*), there exist $\alpha_2, \ldots, \alpha_n \in R$, such that $(a_2 + \alpha_2 a_1)R + \cdots + (a_n + \alpha_n a_1)R = R$. The minimal n satisfying this property is called the *stable range* of R, denoted $\text{sr}(R)$ (so $\text{sr}(R) \leq \dim R + 2$ – strict inequality can hold. Note also that there is "inconsistency up to ± 1" in the literature regarding the definition of the stable range). This property enables one to reduce any $\gamma \in \Gamma$ to $\lambda \in \Lambda$, using a bounded number of elementary operations. Indeed, notice that since all matrices in Γ are invertible, the first row of γ is unimodular. Then, by performing $n - 1$ elementary operations we may create a unimodular $(n-1)$-tuple in the last entries of the first row, and since $1 \in R$, proceed to place 1 in the upper left corner, and use it to annihilate all of the rest of the first row and column. In group theoretic terms, since all the elementary subgroups are conjugate, and any elementary operation is obtained as multiplication by an elementary matrix, this means that Γ is *boundedly generated* by finitely many conjugates of Λ. Like Λ, all of these conjugate fix some point in V_∞, and the Bounded Generation Lemma 2.2 yields a fixed point for Γ, a contradiction which finishes the proof.

It is clear that all that was really relevant to the proof was the stable range of R. Moreover, this notion is similarly defined for *any* ring, not necessarily commutative, only that here one has to distinguish between left and right ideals (although the left and right stable ranges were shown by Vaserstein to be equal [105]). After stating the condition on n in Theorem 1.1 in terms of the stable range in place of dim R, the above proof goes through in this general setting. See [99] for the complete details.

5. Some concluding remarks, questions, and speculations

I. More on Theorem 1.1. The following arises immediately from Theorem 1.1: *Given a finitely generated commutative ring R, when does* $\text{EL}_n(R)$ *begin to have property* (T)? By that theorem the answer lies between $n = 3$ and $n = 2 + \dim R$.

In fact, the proof gives the generally better upper bound $\mathrm{sr}(R)$ (which we suspect is not always optimal). It seems that to address this issue, one should understand better the relation between bounded generation and property (T). We next speculate about a possible strategy.

The proof of Theorem 1.1 actually establishes the following result for any finitely generated ring R and every $n \geq 3$: if $\mathrm{EL}_n(R)$ is boundedly generated by conjugates of $\mathrm{EL}_n(R) \cap \mathrm{GL}_{n-1}(R)$ ($= \Lambda$ in our previous notation), then $\mathrm{EL}_n(R)$ has property (T). The purely algebraic assumption involved here is satisfied when $n \geq \dim R + 2$, and it seems that a full understanding of when it happens (in terms of n as a function of R), takes one beyond the "property (T) territory". However, an attempt at understanding its inverse relation to property (T) should be made: can one show that its failure reflects back on the failure of property (T)? The only device which currently seems available towards such a result, is that of "spaces with walls" defined in [48] (whose more general measurable counterpart is known to capture the lack of property (T) [87]; see also [28]). This setting enables one to construct a negative definite kernel on a group, out of its action on ("half spaces" of) a discrete set, satisfying simple axioms. For our purposes, a natural strategy would thus be to "encode" the algebraic framework into such a set, where the value of the negative definite kernel at $\gamma \in \mathrm{EL}_n(R)$ corresponds to the number of multiplications needed to generate it.

II. Quantifying the robustness of property (T). As mentioned is Section 4.II (2), one of the consequences of the reduced cohomology approach is that any finitely generated Kazhdan group Γ is a quotient of a *finitely presented* Kazhdan group. An intriguing example to which this applies is the group $\Gamma = \mathrm{SL}_3(\mathbb{F}_p[t])$ (which has property (T) because $3 > 2$, and is not finitely presented because $3 < 4$ – see the first mention of the group in this context on [76, p. 134]). Thus, finitely many among the well understood infinite sequence of relations in this group, already suffice to define a Kazhdan group. However, even in this particular case (where explicit Kazhdan constants are known), it is an open problem to *make the existence proof effective*.

The question can in fact be seen as merely one instance of trying to *quantify the robust behavior of property* (T), a phenomenon which was applied in Fisher–Margulis local rigidity results [39]. If Γ has property (T), then for some $\varepsilon > 0$ there is a fixed point for any ε-isometric action of it on a Hilbert space. Moreover, it is actually enough to impose this condition on a generating set only, and for spaces which are Hilbert only locally (on some $1/\varepsilon$ ball). Even further, the action may be well defined only on elements inside a $1/\varepsilon$-ball of Γ, and a moment reflection shows that the latter yields the previously mentioned result about the existence of a finitely presented Kazhdan cover. In fact, one may go further, in assuming the latter action to be only a *near* (ε-)action. In short, by appropriately using a (rather standard by now) limiting argument as needed in the proof of Theorem 1.1, essentially everything in the characterizations of property (T) can be perturbed, yet there is no single non-trivial case when it is known how to do this effectively. To put matters in perspective, we remark that a similar robustness phenomenon *does not* hold for the "opposite" fixed

point property – amenability. Notice also that the proof of Theorem 1.1 *does not yield any explicit Kazhdan constants*. While we believe that some new ideas are needed for the previous questions, it may be that a variation on the proof of Theorem 1.1 will make this an easier task.

III. Property (τ) for irreducible lattices. Although Clozel's recent property (τ) paper [29] largely closes a chapter on the study of property (τ) for arithmetic groups, the story is not quite over as might be perceived. What Clozel actually establishes is the *Selberg property*, and the precise semantics here is important exactly because the congruence subgroup property for many arithmetic groups (e.g., co-compact irreducible lattices in $\mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{R})$) is only conjectured. While Selberg's property, just like his original $\lambda_1 \geq 3/16$ result for $\mathrm{SL}_2(\mathbb{Z})$, yields a spectral gap over the *congruence subgroups*, property (τ) seeks a spectral gap over *all* finite index subgroups. We remark that the incompleteness of property (τ) becomes worse for arithmetic groups over local fields of *positive characteristic*, which were not treated by Clozel.

For higher rank lattices in *simple* algebraic groups, property (T) automatically takes care of property (τ), hence all the unsettled cases lie in a general setting which was extensively and successfully studied in recent years; that of an irreducible lattice Γ in a product $G = G_1 \times G_2$. We believe that under fairly general conditions, certainly ones satisfied when each G_i is a simple algebraic group over a local field, one should be able to get a "softer" proof (certainly avoiding number theory, or a subtle identification of the finite index subgroups), that Γ has property (τ). While as explained, this would have some advantages compared to Clozel's theorem, we note that such a general approach cannot compete with Clozel's: his theorem gives *explicit* spectral bounds on the spectrum of $L^2_0(G/\Gamma_n)$, when restricting to *each simple factor* G_i. Property (τ) is equivalent to the existence of some bound on the joint spectrum of the G_i's.

It may be that there is actually a deeper, more interesting representation theoretic phenomenon underlying the conjectural property (τ) for irreducible lattices: *Is it true that whenever π is a unitary Γ-representation, possibly infinite dimensional, which admits almost invariant vectors, then π has some subrepresentation which extends to G*? Obviously, this would immediately imply property (τ) for Γ (using only that G_i have no non-trivial finite dimensional unitary representations). Such a result would be extremely useful. Some support may be provided by the superrigidity for reduced cohomology in [95], as the failure of property (T) for Γ is always detected by representations with reduced cohomology (Theorem 4.2 above), which in return, *must come from the ambient group G* [95, Theorem 3.1].

IV. Burnside groups and Zelmanov's theorem. The existence of infinite Burnside groups, i.e., finitely generated groups of bounded torsion (first established by Adyan and Novikov, see also [83] and the references therein), is still one of the impressive achievements of infinite group theory. Even harder non-finiteness results are non-amenability of such groups [2]. It is an extremely intriguing question, to which we believe the answer is positive, *whether all Burnside groups should have property* (T). One of the puzzling features of the problem is that it is unclear if it should be attacked

within a *geometric*, or an *algebraic* approach. While serious attempts in the negative have been made within the former direction, we believe that it is actually the latter which should be used, and that reduced cohomology may again become useful. Of course, one of the rewards of establishing property (T) for Burnside groups would be an immediate "size dichotomy" of independent interest: either they are *finite*, or *non-amenable*. There is no known counterexample to this plausible statement.

With the hope of establishing property (T) for Burnside groups in mind, one can wildly speculate further, about a possible approach to reproving Zelmanov's celebrated positive solution to the restricted Burnside problem (cf. [110]): A *residually finite Burnside group is finite*. The idea is to try to adapt Margulis' "(T) ∩ (amenable) = (finite)" strategy in proofs of normal subgroup theorems, to this setting. More precisely, assume that Γ is a Burnside group and $N_i \lhd \Gamma$ is a decreasing sequence of finite index normal subgroups. If Γ was shown to have (T), we would know that Γ/N_i are expanders. On the other hand, a reduction (due to Hall and Higman, already used by Zelmanov) enables one to assume that Γ has prime power torsion, in which case all the finite quotients would have a similar property (and are hence nilpotent). While nilpotency may hint (although in itself is not enough to show) that the quotients should not be expanders, the additional uniform torsion may be sufficient to establish this opposite behavior. Its contrast with the first, property (T) part, would then imply Zelmanov's theorem.

References

[1] Adyan, S. I., Mennicke, J., On bounded generation of $\mathrm{SL}_n(\mathbb{Z})$. *Internat. J. Algebra Comput.* **2** (4) (1992), 357–365.

[2] Adyan, S. I., Random walks on free periodic groups. *Izv. Akad. Nauk SSSR Ser. Mat.* **46** (6) (1982), 1139–1149.

[3] Alon, A., Lubotzky, A., Wigderson, A., Semi-direct product in groups and zig-zag product in graphs: connections and applications (extended abstract). In *42nd IEEE Symposium on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, 2001, 630–637.

[4] Alon, A., Roichman, Y., Random Cayley graphs and expanders. *Random Structures Algorithms* 5 (2) (1994), 271–284.

[5] Babai, L., Seress, A., On the diameter of permutation groups. *European J. Combin.* **13** (4) (1992), 231–243.

[6] Babai, L., Kantor, W. M., Lubotzky, A., Small-diameter Cayley graphs for finite simple groups. *European J. Combin.* **10** (6) (1989), 507–522.

[7] Bader, U., Shalom, Y., Factor and normal subgroup theorems for lattices in products of groups. *Invent. Math.* **163** (2) (2006), 415–454.

[8] Ballmann, W., Świątkowski, J., On L^2-cohomology and property (T) for automorphism groups of polyhedral cell complexes. *Geom. Funct. Anal.* **7** (4) (1997), 615–645.

[9] Bekka, M. B., Kazhdan's property (T) for the unitary group of a separable Hilbert space. *Geom. Funct. Anal.* **13** (3) (2003), 509–520.

[10] Bekka, M. E. B., Cherix, P.-A., Jolissaint, P., Kazhdan constants associated with Laplacian on connected Lie groups. *J. Lie Theory* **8** (1) (1998), 95–110.

[11] Bekka, M. B., de la Harpe, P., Valette, A., *Kazhdan's property* (T). Forthcoming book, 2006.

[12] Bekka, M. E. B., Meyer, M., On Kazhdan's property (T) and Kazhdan constants associated to a Laplacian on $\mathrm{SL}_3(\mathbb{R})$. *J. Lie Theory* **10** (1) (2000), 93–105.

[13] Borel, A., Wallach, N., *Continuous cohomology, discrete subgroups, and representations of reductive groups*. Ann. of Math. Stud. 94, Princeton University Press, Princeton, N.J., 1980.

[14] Bourgain, J., On the Erdös-Volkmann and Katz-Tao ring conjectures. *Geom. Funct. Anal.* **13** (2003), 334–365.

[15] Bourgain, J., Gamburd, A., Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$. Preprint.

[16] Bourgain, J., Gamburd, A., On the spectral gap for finitely-generated subgroups of $\mathrm{SU}(2)$. Preprint.

[17] Bourgain, J., Gamburd, A., Sarnak, P., Sieving and expanders. Preprint, March 2006.

[18] Bourgain, J., Glibichuk, A., Konyagin, S., Estimate for the number of sums and products and for exponential sums in fields of prime order. Preprint.

[19] Bourgain, J., Katz, N., Tao, T., A sum-product estimate in finite fields, and applications. *Geom. Funct. Anal.* **14** (1) (2004), 27–57.

[20] Breuillard, E., Gelander, T., Cheeger constant and algebraic entropy of linear groups. *Internat. Math. Res. Notices* **2005** (56) (2005), 3511–3523.

[21] Burger, M., Kazhdan constants for $\mathrm{SL}_3(\mathbb{Z})$. *J. Reine Angew. Math.* **413** (1991), 36–67.

[22] Burger, M., Mozes, S., Finitely presented simple groups and products of trees. *C. R. Acad. Sci. Paris Sér. I Math.* **324** (7) (1997), 747–752.

[23] Caprace, P. E., Rémy, B., Simplicité abstraite des groupes de Kac-Moody non affines. *C. R. Acad. Sci. Paris Sér. I Math.* **342** (2006), 539–544.

[24] Carter, D., Keller, G., Bounded elementary generation of $\mathrm{SL}_n(\mathcal{O})$. *Amer. J. Math.* **105** (3) (1983), 673–687.

[25] Cartwright, D., Mantero, A. M., Steger, T., Zappa, A., Groups acting simply transitively on the vertices of a building of type $\tilde{A}_2$. II. The cases $q = 2$ and $q = 3$. *Geom. Dedicata* **47** (2) (1993), 167–223.

[26] Cartwright, D. I., Młotkowski, W., Steger, T., Property (T) and $\tilde{A}_2$ groups. *Ann. Inst. Fourier* (*Grenoble*) **44** (1) (1994), 213–248.

[27] Casselman, W., On a p-adic vanishing theorem of Garland. *Bull. Amer. Math. Soc.* **80** (1974), 1001–1004.

[28] Cherix, P. A., Martin, F., Valette, A., Spaces with measured walls, the Haagerup property and property (T). *Ergodic Theory Dynam. Systems* **24** (6) (2004), 1895–1908.

[29] Clozel, L., Démonstration de la conjecture τ. *Invent. Math.* **151** (2) (2003), 297–328.

[30] Colin de Verdière, Y., Spectres de graphes. Cours Spécialisés, Soc. Math. France, Paris 1998.

[31] Corlette, K., Archimedean superrigidity and hyperbolic geometry. *Ann. of Math.* (2) **135** (1) (1992), 165–182.

[32] de Cornulier, Y., Relative Kazhdan Property. *Ann. Sci. École Norm. Sup.* **39** (2) (2006), 301–333.

[33] de Cornulier, Y., Kazhdan property for spaces of continuous functions. *Bull. Belgian Math. Soc.*, to appear.

[34] Davidoff, G., Sarnak, P., Valette, A., Elementary number theory, group theory, and Ramanujan graphs. London Math. Soc. Stud. Texts 55, Cambridge University Press, Cambridge 2003.

[35] Dymara, J., Januszkiewicz, T., New Kazhdan groups. *Geom. Dedicata* **80** (1–3) (2000), 311–317.

[36] Dymara, J., Januszkiewicz, T., Cohomology of buildings and their automorphism groups. *Invent. Math.* **150** (3) (2002), 579–627.

[37] Erdös, E., Szemerédi, P., On sums and products of integers. In *Studies in pure mathematics*, Birkhäuser, Basel 1983, 213–218.

[38] Eskin, A., Mozes, S., Oh, H., On uniform exponential growth for linear groups. *Invent. Math.* **160** (1) (2005), 1–30.

[39] Fisher, D., Margulis, G. A., Almost isometric actions, property (T), and local rigidity. *Invent. Math.* **162** (2005), 19–80.

[40] Gamburd, A., On the spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbb{Z})$. *Israel J. Math.* **127** (2002), 157–200.

[41] Gamburd, A., Hoory, S., Shahshahani, M., Shalev, A., Virag, B., On the girth of random Cayley graphs. Preprint, 2005.

[42] Gamburd, A., Jakobson, D., and Sarnak, P., Spectra of elements in the group ring of $\mathrm{SU}(2)$. *J. Eur. Math. Soc.* **1** (1) (1999), 51–85.

[43] Gamburd, A., Mehrdad, S., Uniform diameter bounds for some families of Cayley graphs. *Internat. Math. Res. Notices* **2004** (71) (2004), 3813–3824.

[44] Garland, H., p-adic curvature and the cohomology of discrete subgroups of p-adic groups. *Ann. of Math.* (2) **97** (1973), 375–423.

[45] Ghys, E., Groupes aléatoires (d'après Misha Gromov,...). *Astérisque* **294** (2004), 173–204.

[46] Gromov, M., Hyperbolic groups. In *Essays in group theory*, Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–263.

[47] Gromov, M., Random walk in random groups. *Geom. Funct. Anal.* **13** (1) (2003), 73–146.

[48] Haglund, F., Paulin, F., Simplicité de groupes d'automorphismes d'espaces à courbure négative. In *The Epstein birthday schrift*, Geom. Top. Mon. 1, Geom. Topol. Publ., Coventry 1998, 181–248 (electronic).

[49] Hahn, A. J., O'Meara, O. T., *The classical groups and K-theory*. Grundlehren Math. Wiss. 291, Springer-Verlag, Berlin 1989.

[50] de la Harpe, P., Valette, A., La propriété (T) de Kazhdan pour les groupes localement compacts (avec un appendice de Marc Burger). *Astérisque* **175**, 1989.

[51] Helfgott, H., Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Preprint, 2005.

[52] Hrushovski, E., Pillay, A., Definable subgroups of algebraic groups over finite fields. *J. Reine Angew. Math.* **462** (1995), 69–91.

[53] Izeki, H., Nayatani, S., Combinatorial harmonic maps and discrete-group actions on Hadamard spaces. *Geom. Dedicata* **114** (2005), 147–188.

[54] Jolissaint, P., On property (T) for pairs of topological groups. *Enseign. Math.* (2) **51** (1–2) (2005), 31–45.

[55] van der Kallen, W., $SL_3(\mathbf{C}[X])$ does not have bounded word length. In *Algebraic K-theory* (Oberwolfach, 1980), Part I, Lecture Notes in Math. 966, Springer-Verlag, Berlin 1982, 357–361.

[56] Kassabov, M., Kazhdan constants for $SL_n(\mathbb{Z})$. *Internat. J. Algebra Comput.* **15** (5–6) (2005), 971–995.

[57] Kassabov, M., Symmetric groups and expander graphs. Preprint; arXiv:math.GR/0505624.

[58] Kassabov, M., Universal lattices and unbounded rank expanders. Preprint; arXiv:math.GR/0502237.

[59] Kassabov, M., Symmetric groups and expanders. *Electron. Res. Announc. Amer. Math. Soc.* **11** (2005), 47–56 (electronic).

[60] Kassabov, M., Lubotzky, A., Nikolov, N., Finite simple groups as expanders. *Proc. Nat. Acad. Sci.* **103** (16) (2006), 6116–6119.

[61] Kassabov, M., Nikolov, N., Universal lattices and property (τ). *Invent. Math.* **165** (1) (2006), 209–224.

[62] Kazhdan, D. A., On the connection of the dual space of a group with the structure of its closed subgroups. *Funk. Anal. Pril.* **1** (1967), 71–74.

[63] Konyagin, S. V., A sum-product estimate in fields of prime order. Preprint.

[64] Kumaresan, S., On the canonical K-types in the irreducible unitary g-modules with nonzero relative cohomology. *Invent. Math.* **59** (1) (1980), 1–11.

[65] Lafferty, J. D., Rockmore, D., Fast Fourier analysis for SL_2 over a finite field and related numerical experiments. *Exper. Math.* **1** (2) (1992), 115–139.

[66] Lafferty, J. D., Rockmore, D., Numerical investigation of the spectrum for certain families of Cayley graphs. In *Expanding graphs*, DIMACS Ser. Disc. Math. Theo. Comput. Sci. 10, Amer. Math. Soc., Providence, RI, 1993, 63–73.

[67] Lafferty, J. D., Rockmore, D., Level spacings for Cayley graphs. In *Emerging applications of number theory*, IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 373–386.

[68] Louvet, N., À propos d'un théorme de Vershik et Karpushev. *Enseign. Math.* (2) **47** (2001), 287–314.

[69] Lubotzky, A., *Discrete groups, expanding graphs and invariant measures*. With an appendix by J. D. Rogawski, Progr. Math. 125, Birkhäuser, Basel 1994.

[70] Lubotzky, A., Cayley graphs: eigenvalues, expanders and random walks. In *Surveys in combinatorics*, London Math. Soc. Lecture Note Ser. 218, Cambridge University Press, Cambridge 1995, 155–189.

[71] Lubotzky, A., Finite simple groups of Lie type as expanders. In preparation.

[72] Lubotzky, A., Samuels, B., Vishne, U., Explicit constructions of Ramanujan complexes of type A_d. *European J. Combin.* **26** (6) (2005), 965–993.

[73] Lubotzky, A., Weiss, B., Groups and expanders. In *Expanding graphs*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 10, Amer. Math. Soc., Providence, RI, 1993, 95–109.

[74] Margulis, G. A., Explicit constructions of expanders. *Prob. Pered. Inform.* **9** (4) (1973), 71–80.

[75] Margulis, G. A., Finitely-additive invariant measures on Euclidean spaces. *Ergodic Theory Dynam. Systems* **2** (3–4) (1982), 383–396.

[76] Margulis, G. A., *Discrete subgroups of semisimple Lie groups*. Ergeb. Math. Grenzgeb. (3) 17, Springer-Verlag, Berlin 1991.

[77] Martin, F., Reduced 1-cohomology of connected locally compact groups and applications, *J. Lie Theory* **16** (2006), 311–328.

[78] Neuhauser, M., Kazhdan's property T for the symplectic group over a ring. *Bull. Belg. Math. Soc.* **10** (4) (2003), 537–550.

[79] Nikolov, N., A product decomposition for the classical quasisimple group. *J. Lie Theory*, to appear.

[80] Oh, H., Uniform pointwise bounds for matrix coefficients of unitary representations and applications to Kazhdan constants. *Duke Math. J.* **113** (1) (2002), 133–192.

[81] Ollivier, Y., A January 2005 invitation to random groups. Preprint.

[82] Ollivier, Y., Spectral interpretations of property (T). Preprint.

[83] Olshanski, A. Y., The Novikov-Adyan theorem. *Mat. Sb.* (*N.S.*) **118** (160) (2) (1982), 203–235.

[84] Pansu, P., Formules de Matsushima, de Garland et propriété (T) pour des groupes agissant sur des espaces symétriques ou des immeubles. *Bull. Soc. Math. France* **126** (1) (1998), 107–139.

[85] Platonov, V., Rapinchuk, A., *Algebraic groups and number theory*, Pure and Applied Mathematics 139, Academic Press Inc., Boston, MA, 1994.

[86] Reingold, O., Vadhan, S., Wigderson, A., Entropy waves, the zig-zag graph product, and new constant-degree expanders. *Ann. of Math.* (2) **155** (1) (2002), 157–187.

[87] Robertson, G., Steger, T., Negative definite kernels and a dynamical characterization of property (T) for countable groups. *Ergodic Theory Dynam. Systems* **18** (1) (1998), 247–253.

[88] Roichman, Y., Upper bound on the characters of the symmetric groups. *Invent. Math.* **125** (3) (1996), 451–485.

[89] Roichman, Y., Expansion properties of Cayley graphs of the alternating groups. *J. Combin. Theory Ser. A* **79** (2) (1997), 281–297.

[90] Sarnak, P., *Some applications of modular forms*. Cambridge Tracts in Math. 99, Cambridge University Press, Cambridge 1990.

[91] Sarnak, P., Xue, X., Bounds for multiplicities of automorphic representations. *Duke Math. J.* **64** (1991), 207–227.

[92] Shalom, Y., Expanding graphs and invariant means. *Combinatorica* **17** (4) (1997), 555–575.

[93] Shalom, Y., Expander graphs and amenable quotients. In *Emerging applications of number theory*, IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 571–581.

[94] Shalom, Y., Explicit Kazhdan constants for representations of semisimple and arithmetic groups. *Ann. Inst. Fourier* (*Grenoble*) **50** (3) (2000), 833–863.

[95] Shalom, Y., Rigidity of commensurators and irreducible lattices. *Invent. Math.* **141** (1) (2000), 1–54.

[96] Shalom, Y., Bounded generation and Kazhdan's property (T). *Inst. Hautes Études Sci. Publ. Math.* **90** (1999), 145–168.

[97] Shalom, Y., Rigidity, unitary representations of semisimple groups, and fundamental groups of manifolds with rank one transformation group. *Ann. of Math.* (2) **152** (1) (2000), 113–182.

[98] Shalom, Y., Harmonic analysis, cohomology, and the large-scale geometry of amenable groups. *Acta Math.* **192** (2) (2004), 119–185.

[99] Shalom, Y., Elementary linear groups and Kazhdan's property (T). In preparation.

[100] Silberman, L., Addendum to "Random walk in random groups" by M. Gromov. *Geom. Funct. Anal.* **13** (1) (2003), 147–177.

[101] Suslin, A. A., The structure of the special linear group over rings of polynomials. *Izv. Akad. Nauk SSSR Ser. Mat.* **41** (2) (1977), 235–252 (in Russian).

[102] Tao, T., Non-commutative sum set estimates. Preprint.

[103] Tao, T., Vu, V., *Additive combinatorics*. Cambridge Stud. Adv. Math. 105, Cambridge University Press, to appear, 2006.

[104] Valette, A., Nouvelles approches de la propriété (T) de Kazhdan. *Astérisque* **294** (2004), 97–124.

[105] Vaserstein, L. N., The stable range of rings and the dimension of topological spaces. *Funk. Anal. Pril.* **5** (2) (1971), 17–27 (in Russian).

[106] Vershik, A., Karpushev, S., Cohomology of groups in unitary representations, neighborhood of the identity and conditionally positive definite functions. *Mat. Sb.* (*N.S.*) **119** (4) (1982), 521–533 (in Russian).

[107] Vogan, D. A., Zuckerman, G. J., Unitary representations with nonzero cohomology. *Compositio Math.* **53** (1) (1984), 51–90.

[108] Wang, M. T., Generalized harmonic maps and representations of discrete groups. *Comm. Anal. Geom.* **8** (3) (2000), 545–563.

[109] Witte, D., Bounded generation of SL(n, A) (after D. Carter, G. Keller, and E. Paige). Preprint, 2005; arXiv:math.GR/0503083.

[110] Zelmanov, E., On the restricted Burnside problem. *Proceedings of the International Congress of Mathematicians* (Kyoto, 1990), Vol. I, The Mathematical Society of Japan, Tokyo, Springer-Verlag, Tokyo, 1991, 395–402.

[111] Żuk, A., La propriété (T) de Kazhdan pour les groupes agissant sur les polyèdres. *C. R. Acad. Sci. Paris Sér. I Math.* **323** (5) (1996), 453–458.

[112] Żuk, A., Property (T) and Kazhdan constants for discrete groups. *Geom. Funct. Anal.* **13** (3) (2003), 643–670.

School of Mathematical Sciences, Tel-Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel
E-mail: yeshalom@post.tau.ac.il

Rankin–Selberg integrals, the descent method, and Langlands functoriality

David Soudry

Abstract. In this article I survey the descent method of Ginzburg, Rallis and Soudry and its main applications to the Langlands functorial lift of automorphic, cuspidal, generic representations on a classical group to (appropriate) GL_n, and to establishing a local Langlands reciprocity law for (split) SO_{2n+1} (joint work with D. Jiang). The descent method arises when we consider certain residues of special cases of a family of global integrals, attached to pairs of automorphic, cuspidal representations, one on a classical group G and one on GL_n. The last part of this article focuses on the case $G = \mathrm{SO}_m$ (split), and the progress made in a joint work with S. Rallis, towards establishing, via the converse theorem, the functorial lift from any automorphic, cuspidal representation on G to $\mathrm{GL}_{2[\frac{m}{2}]}$.

Mathematics Subject Classification (2000). Primary 11F70; Secondary 11R39.

Keywords. L-functions, functorial lift, descent method, Gelfand–Graev models.

1. Introduction

Let F be a number field, and let $\mathbb{A}$ be its ring of Adeles. I will start with a general family of global integrals of Rankin–Selberg type or of Shimura type for $G \times \mathrm{GL}_n$, where G is an orthogonal group, a unitary group, a symplectic group, or a metaplectic group defined over F. This family contains, at one end, the integrals studied by Shahidi, giving rise to the Langlands–Shahidi method, and at another end, this family contains the global integrals, giving rise to the descent method of Ginzburg, Rallis and Soudry. In this section, I sketch the structure of these integrals, and in the next section I report on my joint work with D. Ginzburg and S. Rallis on the descent method and its many applications, and on my joint work with D. Jiang on the local Langlands reciprocity law for the split group SO_{2n+1}. In the third section I report on a joint work in progress with S. Rallis towards the (weak) functorial lift from SO_m to $\mathrm{GL}_{2[\frac{m}{2}]}$.

a. Let E be either F or a quadratic extension of F. Let V and V' be vector spaces over E of dimensions m and m' equipped with non-degenerate bilinear forms b and b', respectively, which are symmetric if $E = F$ and Hermitian (with respect to the Galois conjugation "-" of E over F) otherwise. Let $G = U(V)$ and $G' = U(V')$ be the isometry groups of (V, b) and (V', b'), respectively. Denote by X_i the (orthogonal) direct sum of i hyperbolic planes, where, in the second case, we mean two

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

dimensional spaces over E with Hermitian form (according to some basis) given by $b((x_1, x_2), (y_1, y_2)) = x_1\bar{y}_2 + x_2\bar{y}_1$. Fix a decomposition $X_n = X_n^+ + X_n^-$, where $X_n^{\pm}$ are transversal n-dimensional totally isotropic subspaces. Assume that m and m' have different parities and that one of the following holds.

(1) There is an orthogonal decomposition of the form $V = X_j \oplus Y$ and there is a non-isotropic vector $y_0 \in Y$ such that $V' = Y \cap y_0^{\perp}$, and b' is the restriction of b to $V' \times V'$.

(2) The same as in (1), but reversing the roles of V and V', i.e. $V' = X_j \oplus Y'$ and there is a non-isotropic $y_0' \in Y'$ such that $V = Y' \cap y_0'^{\perp}$, and b is the restriction of b' to $V \times V$.

Let $W = X_n \oplus V'$ and $H = U(W)$. Let π, ρ, τ be irreducible, automorphic, cuspidal representations of $G_{\mathbb{A}}$, $G'_{\mathbb{A}}$, $\mathrm{GL}_n(\mathbb{A}_E)$, respectively. Let P be the parabolic subgroup of H which preserves X_n^+; its Levi part M is isomorphic to $\mathrm{Res}_{E/F}\,\mathrm{GL}_n \times G'$. Thus we may view $\tau|\det|^{s-\frac{1}{2}} \otimes \rho = \tau_s \otimes \rho$ as a representation of $M_{\mathbb{A}}$ and consider its parabolic induction $I(\tau_s, \rho)$ to $H_{\mathbb{A}}$. Let $E(h, f_{\tau_s,\rho})$ be an Eisenstein series corresponding to an analytic section $f_{\tau_s,\rho}$ in $I(\tau_s, \rho)$. Now we distinguish three cases.

(i) $m' < m < 2n + m'$.

(ii) $2n + m' < m$.

(iii) $m < m'$.

In cases (i) and (iii) we apply to $E(h, f_{\tau_s,\rho})$ a Fourier coefficient of Gelfand–Graev type stabilized by G; let us denote it by $E^{\psi,G}(h, f_{\tau_s,\rho})$, where ψ is a non-trivial character of $F\backslash\mathbb{A}$. We pair this coefficient with cusp forms φ_π in π,

$$\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho}) = \int_{G_F\backslash G_{\mathbb{A}}} \varphi_\pi(g) E^{\psi,G}(g, f_{\tau_s,\rho})\, dg. \tag{1.1}$$

In case (ii) we apply to φ_π a Fourier coefficient of Gelfand–Graev type stabilized by H, and pair it with $E(h, f_{\tau_s,\rho})$,

$$\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho}) = \int_{H_F\backslash H_{\mathbb{A}}} \varphi_\pi^{\psi,H}(h) E(h, f_{\tau_s,\rho})\, dh. \tag{1.2}$$

For the general notion of a Fourier coefficient of Gelfand–Graev type see [29]. The integrals above, in this generality, appear in [9] for orthogonal groups. See also [34]. These integrals are meromorphic in the whole plane, and their poles are included in the set of poles of the Eisenstein series involved. They can be unwinded for $\mathrm{Re}(s) \gg 0$ and can be shown to depend, through an inner integration, on an invariant bilinear pairing $b_{\pi,\rho}(\varphi_\pi, \varphi_\rho)$, where, in cases (i) and (ii),

$$b_{\pi,\rho}(\varphi_\pi, \varphi_\rho) = \int_{G'_F\backslash G'_{\mathbb{A}}} \varphi_\pi^{\psi,G'}(g)\varphi_\rho(g)\, dg, \tag{1.3}$$

and, in case (iii),

$$b_{\pi,\rho}(\varphi_\pi, \varphi_\rho) = \int_{H_F \backslash H_{\mathbb{A}}} \varphi_\pi(h) \varphi_\rho^{\psi, G}(h)\, dh. \tag{1.4}$$

In particular, if $b_{\pi,\rho}(\varphi_\pi, \varphi_\rho) = 0$ (identically), then the integrals (1.1), (1.2) are (identically) zero. The integrals above have easy variants when G, G' are special orthogonal groups.

b. Let now (V, b), (V', b') be symplectic spaces over F, and let X_j be a symplectic space of dimension $2j$ over F with two transversal Lagrangians $X_j^{\pm}$. Let $W = X_n \oplus V'$. Let G, G', H be the symplectic groups of V, V', W, respectively. Let P be the parabolic subgroup of H which preserves X_n^+. Denote by $\tilde{H}_{\mathbb{A}}$ the metaplectic cover of $H_{\mathbb{A}}$, and similarly for G, G'. Let π, ρ, τ be irreducible, automorphic, cuspidal representations of $G_{\mathbb{A}}$, $\tilde{G}'_{\mathbb{A}}$, $\mathrm{GL}_n(\mathbb{A})$, respectively. We assume that ρ is genuine. Consider the Eisenstein series $E(h, f_{\tau_s,\rho})$ on $\tilde{H}_{\mathbb{A}}$ corresponding to parabolic induction from $\tau_s \otimes \rho$ (we have to multiply τ_s by a Weil factor corresponding to ψ). Now we can either apply a Fourier–Jacobi coefficient, stabilized by G, to the Eisenstein series and pair it, as above, along $G_F \backslash G_{\mathbb{A}}$ with cusp forms φ_π, or apply such a coefficient, stabilized by H, to φ_π and pair it along $H_F \backslash H_{\mathbb{A}}$ with the Eisenstein series. In this way we get global integrals $\mathcal{L}(\varphi_\pi, \phi, f_{\tau_s,\rho})$; here ϕ is a Schwartz function in the space of the Weil representation, which occurs in the Fourier–Jacobi coefficients above. As before these integrals are meromorphic, can be unwinded, and depend through inner integrations on invariant bilinear pairings, which define a Fourier–Jacobi model of π with respect to ρ or vice-versa. Similar integrals can be written for π on $\tilde{G}_{\mathbb{A}}$, ρ on $G'_{\mathbb{A}}$, and τ as before (with $H_{\mathbb{A}}$ instead of $\tilde{H}_{\mathbb{A}}$). Finally, similar integrals can also be written when (V, b), (V', b') are Hermitian spaces over F. For more details see [29], [15], [34].

c. Two extreme cases

1. Assume that G is trivial. Also assume that when b' is symmetric, then G' is a special orthogonal group. Then, in all cases above, one can see that G' must be quasi-split, and the global integrals above are nothing but applications of a Whittaker coefficient to the Eisenstein series $E(h, f_{\tau_s,\rho})$, and using (1.4) and its analog in (b) above, we see that (for the global integrals to be non-trivial) ρ must be globally generic. This is the well-known Langlands–Shahidi method. It is worked out in a long series of papers (see [7] for a survey) by Shahidi and constitutes a beautiful chapter in mathematics. In particular, he established the complete theory of the standard L-functions $L(\rho \times \tau, s)$ (except the metaplectic case) and of $L(\tau, \wedge^2, s)$, $L(\tau, \mathrm{sym}^2, s)$, $L(\tau, \mathrm{Asai}, s)$. Together with the converse theorem of Cogdell and Piatetski–Shapiro [3] it yields the existence of the weak functorial lift from cuspidal generic representations of $G'_{\mathbb{A}}$ to automorphic representations of $\mathrm{GL}_N(\mathbb{A}_E)$ (appropriate N) [1], [2], [26].

2. Assume that G' is trivial. Also assume that when b is symmetric, then G is a special orthogonal group. Then the Levi part of the parabolic subgroup P is isomorphic to $\mathrm{Res}_{E/F}\,\mathrm{GL}_n$. The global integrals, in this case, were studied by many authors; see, for example, [8], [15], [6], [34], [35], [36]. As before, using (1.3) and its analog in (b) above, we see that π must be generic (or else, the global integrals are trivial). These integrals yield (up to a controllable factor) the quotient of the partial L-function $L^S(\pi \times \tau, s)$ by the following denominator. It is $L^S(\tau, r_G, 2s)$, by which we mean $L^S(\tau, \wedge^2, 2s)$ (G odd orthogonal), or $L^S(\tau, \mathrm{sym}^2, 2s)$ (G even orthogonal or symplectic), or $L^S(\tau, \mathrm{Asai}, 2s)$ (G odd unitary); it is $L^S\big(\tau, s+\frac{1}{2}\big)L^S(\tau, \wedge^2, 2s)$ if π is on a metaplectic group, and finally, if G is even unitary, the denominator is the partial Asai L-function of τ at $2s$, but in this case the numerator is $L^S(\pi \times \tau\gamma^{-1}, s)$, where we twist τ by a character γ^{-1}, which enters in the choice of the Weil representation defining the Fourier–Jacobi coefficient in this case. We remark that in the metaplectic case $L^S(\pi \times \tau, s)$ depends also on ψ. The descent method of Ginzburg, Rallis and Soudry is derived when we analyze the existence of a pole at $s = 1$ for the global integrals $\mathcal{L}(\varphi_\pi, f_{\tau_s})$, $\mathcal{L}(\varphi_\pi, \phi, f_{\tau_s})$. The descent method allows us to give a complete description of the image of the weak functorial lift from cuspidal generic representations of $G_{\mathbb{A}}$ to $\mathrm{GL}_N(\mathbb{A}_E)$, to prove the existence and give a description of endoscopic lifts to G (from generic cuspidal representations), to obtain a full local Langlands reciprocity law for generic representations of SO_{2n+1} (joint work with D. Jiang), and much more. I survey some of these applications in the next section.

The general case (G, G' non-trivial) was considered mainly for orthogonal groups in [9], where it is shown that the integrals $\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho})$ yield the quotient

$$\frac{L^S(\pi \times \tau, s)}{L^S\big(\rho \times \tau, s+\frac{1}{2}\big)L^S(\tau, r, 2s)},$$

where $r = r_G = \wedge^2$ or sym^2, depending on G. See also [4] for local analogs in case $m < m'$.

In the third section, I report on the progress in a joint work with S. Rallis towards the existence of a weak functorial lift of cuspidal (not necessarily generic) representations of $\mathrm{SO}_m(\mathbb{A})$ to $\mathrm{GL}_{2[\frac{m}{2}]}(\mathbb{A})$.

2. The descent method and applications

We retain the notation from part c.2 of the previous section. So we have to assume that π is (globally) generic. We assume for simplicity that if G is even (special) orthogonal, then it is split. We also assume, as we may, that τ is unitary such that its central character ω_τ is trivial on the positive real numbers, embedded diagonally at all archimedean primes in the Idele group. The material in Section 2.1–2.3 is part of a long-term joint work with Ginzburg and Rallis [10]–[14], [34].

2.1.

Theorem 2.1. *$L^S(\pi \times \tau, s)$ is holomorphic for* $\mathrm{Re}(s) > 1$ *except when $n = 1$, τ is trivial, and π is on a metaplectic group; in this case the only such pole may occur at $s = \frac{3}{2}$. If $L^S(\pi \times \tau, s)$ has a pole at s_0 with* $\mathrm{Re}(s_0) = 1$, *then $s_0 = 1$. In this case $L^S(\tau, r_G, s)$ has a pole at $s = 1$, when G is orthogonal, symplectic, or odd unitary; if π is on a metaplectic group, then $L^S(\tau, \wedge^2, s)$ has a pole at $s = 1$ and $L\left(\tau, \frac{1}{2}\right) \neq 0$. Finally, if G is even unitary, then $L^S(\hat{\tau} \times \gamma, \mathrm{Asai}, s)$ has a pole at $s = 1$.*

This follows from the fact that the analysis of poles of the integrals $\mathcal{L}(\varphi_\pi, f_{\tau_s})$, $\mathcal{L}(\varphi_\pi, \phi, f_{\tau_s})$ reduces to that of the Eisenstein series $E(h, f_{\tau_s})$ induced from τ.

In a similar way we get that if τ is the (standard) weak functorial lift of π, then τ is self-conjugate, and its central character is trivial on $\mathbb{A}^* = \mathbb{A}_F^*$, and the results of Theorem 2.1 hold for τ. The reason for this is that $L^S(\pi \times \hat{\tau}, s)$ has a pole at $s = 1$. Moreover, for the residue of the global integrals at $s = 1$ to be non-trivial, the L^2-pairing between π and $\mathrm{Span}\{\mathrm{Res}_{s=1} E^{\psi^{-1}, G}(\cdot, f_{\tilde{\tau}_s \gamma})_{|G_{\mathbb{A}}}\} := \sigma_\psi(\tau)$ is non-trivial; in case π is on a metaplectic group, then, in the definition of $\sigma_\psi(\tau)$, we have to take restrictions to $\tilde{G}_{\mathbb{A}}$. Here γ is trivial except when G is even unitary. Note also that starting with τ, G (resp. $\tilde{G}$) is determined by n and the precise data about the pole at $s = 1$ related to τ. We keep all this implicit in the notation $\sigma_\psi(\tau)$. One of our main theorems is

Theorem 2.2. *Let τ be an irreducible, automorphic, cuspidal representation of* $\mathrm{GL}_n(\mathbb{A}_E)$ *such that ω_τ is trivial on $\mathbb{A}^*$. Assume that the results of Theorem* 2.1 *about the pole at $s = 1$ are satisfied for τ. Then $\sigma_\psi(\tau)$ is a non-trivial, automorphic, cuspidal, multiplicity free representation of $G_{\mathbb{A}}$ (respectively of $\tilde{G}_{\mathbb{A}}$ if H is symplectic). All irreducible summands of $\sigma_\psi(\tau)$ are (ψ-)generic and lift at almost all finite places to τ. Each such representation has a non-trivial L^2-pairing with $\sigma_\psi(\tau)$.*

2.2. We call $\sigma_\psi(\tau)$, for τ as in the last theorem, the descent of τ to G (resp. to $\tilde{G}$). Theorem 2.2 describes the cuspidal part of the weak functorial lift from generic cuspidal representations on G (resp. $\tilde{G}$) to GL_n, without knowing that such a lift exists in the sense that we know which cuspidal τ can occur in the image, and moreover, we do construct in a direct manner for such τ irreducible, cuspidal, (ψ-)generic representations which lift to τ; these are the summands of $\sigma_\psi(\tau)$.

When we analyze non-cuspidal τ on $\mathrm{GL}_n(\mathbb{A})$ which may be obtained as a weak lift from cuspidal generic representations of $G_{\mathbb{A}}$ (resp. a metaplectic group), we get that the central character of τ is trivial on $\mathbb{A}^*$, and, by successive applications of Theorem 2.1, we get that, except in the metaplectic case, τ must have the form $\tau = \tau_1 \times \cdots \times \tau_l$, where the τ_i are pairwise inequivalent, irreducible, unitary, self-conjugate automorphic representations of $\mathrm{GL}_{n_i}(\mathbb{A}_E)$, cuspidal when $n_i > 1$; each one satisfying the results of Theorem 2.1 about the pole at $s = 1$. In the metaplectic case τ may also have the form above, with an added "tail" of the form $|\cdot|^{\frac{1}{2}} \times |\cdot|^{-\frac{1}{2}}$.

Thus, for such τ, except in the last case, we form $\sigma_\psi(\tau)$ as before by replacing the residue of the Eisenstein series at $s = 1$, with the multi-residue at $(1, \dots, 1)$ of the Eisenstein series induced from $\tau_{1,s_1} \times \cdots \times \tau_{l,s_l}$. We prove

Theorem 2.3. *Let τ be an irreducible automorphic representation of $\mathrm{GL}_n(\mathbb{A}_E)$ as in the last paragraph (except the additional possibility in the metaplectic case). Then $\sigma_\psi(\tau)$ satisfies the conclusions of Theorem 2.2.*

Again, the descent $\sigma_\psi(\tau)$ of τ constructs cuspidal, generic representations of $G_\mathbb{A}$ (resp. $\tilde{G}_\mathbb{A}$) which lift to τ. Since all cuspidal, generic representations of $G_\mathbb{A}$ do lift to $\mathrm{GL}_n(\mathbb{A}_E)$, by [2] (recall again that n and G are related) we get the description of the image of this lift and that the descent is an explicit inverse map of this lift to the set of near equivalence classes of irreducible, cuspidal, generic representations of $G_\mathbb{A}$. Moreover, since each factor τ_i of τ, as above, satisfies the assumptions of Theorem 2.2, it determines a corresponding group G_i (or $\tilde{G}_i$) and a cuspidal, generic representation on it, π_i, which lifts to τ_i. Thus we establish the endoscopic lift from cuspidal, generic representations on $G_{1,\mathbb{A}} \times \cdots \times G_{l,\mathbb{A}}$ to automorphic representations on $G_\mathbb{A}$. In particular, if τ is non-cuspidal and in the image of the weak lift above, say, lifted from π (cuspidal, generic), then π is in the image of an endoscopic lift which can be described precisely.

Example. Let $G = \mathrm{Sp}_{2k}$ or SO_{2k} (split). In the first case $n = 2k + 1$, and in the second case $n = 2k$. Let π be an irreducible, automorphic, cuspidal, generic representation of $G_\mathbb{A}$. Assume that the lift τ of π to $\mathrm{GL}_n(\mathbb{A})$ is non-cuspidal. Then $\tau = \tau_1 \times \cdots \times \tau_l$, as above. Assume, for simplicity, that all $\omega_{\tau_i} = 1$. Then all partial L-functions $L^S(\tau_i, \mathrm{sym}^2, s)$ have a pole at $s = 1$. Let $G_i = \mathrm{Sp}_{2k_i}$ if $n_i = 2k_i + 1$ is odd, and $G_i = \mathrm{SO}_{2k_i}$ (split) if $n_i = 2k_i$ is even; $n = n_1 + \cdots + n_l$. Let π_i be an irreducible summand of the descent $\sigma_\psi(\tau_i)$ to G_i. Then π is a weak (generalized endoscopic) lift of $\pi_1 \otimes \cdots \otimes \pi_l$ from $G_1 \times \cdots \times G_l$.

2.3. The descent has a local analog, which was developed in [11], [12] for the target group $\tilde{G} = \tilde{\mathrm{Sp}}_{2k}$. We proved

Theorem 2.4. *Let K be a local non-archimedean field of characteristic zero. Let $\tau_1, \dots, \tau_l$ be pairwise inequivalent, irreducible, supercuspidal representations of $\mathrm{GL}_{2k_1}(K), \dots, \mathrm{GL}_{2k_l}(K)$, respectively, such that each local L-function $L(\tau_i, \wedge^2, s)$ has a pole at $s = 0$. Let $\tau = \tau_1 \times \cdots \times \tau_l$ be the corresponding parabolic induction to $\mathrm{GL}_{2k}(K)$, where $k = k_1 + \cdots + k_l$. Then there is a unique (up to isomorphism) irreducible, supercuspidal, ψ-generic representation π of $\tilde{\mathrm{Sp}}_{2k}(K)$ such that the local gamma factor $\gamma(\pi \times \tau, s, \psi)$ has a pole of order l at $s = 1$. The representation π is obtained by a local analogue, applied to τ, of the descent construction.*

For the local analogue of the descent construction, we induce, as in the global set-up, $\tau = \tau_{1,s_1} \times \cdots \times \tau_{l,s_l}$ at the point $s_1 = \cdots = s_l = 1$ from the Siegel parabolic subgroup P to $H = \mathrm{Sp}_{4k}(K)$, and we consider the corresponding Langlands quotient,

call it e_τ; this is the analogue to the multi-residue at $(1, \dots 1)$ of the Eisenstein series in the global case. Then we apply to e_τ a Jacquet module, analogous to the Fourier–Jacobi coefficient, that we apply in the global case to the multi-residue of the Eisenstein series.

2.4. The local descent from $\mathrm{GL}_{2k}(K)$ to $\mathrm{SO}_{2k+1}(K)$ yields powerful results, among which are the local converse theorem for generic representations of $\mathrm{SO}_{2k+1}(K)$, a full local Langlands reciprocity law for generic representations of $\mathrm{SO}_{2k+1}(K)$, a rigidity property (strong multiplicity one, up to isomorphism) of irreducible, cuspidal generic representations of $\mathrm{SO}_{2k+1}(\mathbb{A})$, and more.... These are results of my joint work with D. Jiang and can all be found in [23], [24]. I survey this work in this subsection.

Consider the representations τ and π as in Theorem 2.4. Using the local Howe duality from $\tilde{\mathrm{Sp}}_{2k}(K)$ to $\mathrm{SO}_{2k+1}(K)$, we lift π to an irreducible, supercuspidal, generic representation σ of $\mathrm{SO}_{2k+1}(K)$. It is unique, up to isomorphism, with the property that the local gamma factor $\gamma(\sigma \times \tau, s, \psi)$ has a pole of order l at $s = 1$. Using the existence of the weak lift of [1] we prove

Theorem 2.5 (The local converse theorem). *Let σ and σ' be two irreducible generic representations of $\mathrm{SO}_{2k+1}(K)$ such that, for all $j < 2k$ and all irreducible generic representations ρ of $\mathrm{GL}_j(K)$,*

$$\gamma(\sigma \times \rho, s, \psi) = \gamma(\sigma' \times \rho, s, \psi).$$

Then σ and σ' are isomorphic.

The idea is to reduce the proof to supercuspidal representations σ and σ', lift them locally to $\mathrm{GL}_{2k}(K)$, and use Henniart's local converse theorem for $\mathrm{GL}_n(K)$ [18] to conclude that both representations lift to the same representation τ of $\mathrm{GL}_{2k}(K)$. We prove that τ has the form as in Theorem 2.4, and then we conclude that both gamma factors of σ, twisted by τ, and of σ', twisted by τ, have a pole of order l at $s = 1$. Using the uniqueness mentioned in the last paragraph, we conclude that σ and σ' are isomorphic. As a result from this we prove

Theorem 2.6. *There is a one-to-one correspondence t between the isomorphism classes of irreducible, supercuspidal, generic representations of $\mathrm{SO}_{2k+1}(K)$ and the isomorphism classes of irreducible, generic representations of $\mathrm{GL}_{2k}(K)$, as in Theorem 2.4, such that if $\tau = t(\sigma)$, then for all irreducible generic representations ρ of $\mathrm{GL}_j(K)$, $j > 0$,*

$$\begin{aligned} \gamma(\sigma \times \rho, s, \psi) &= \gamma(\tau \times \rho, s, \psi), \\ L(\sigma \times \rho, s) &= L(\tau \times \rho, s). \end{aligned} \tag{2.1}$$

See [2] for partial results for the other groups G. Let us outline some applications.

Theorem 2.7 (Rigidity theorem). *Let σ and σ' be two irreducible, automorphic, cuspidal, generic representations of $\mathrm{SO}_{2k+1}(\mathbb{A})$ which are isomorphic, at almost all*

unramified places. Then σ and σ' are isomorphic. In particular, the weak lift from cuspidal, generic representations of $\mathrm{SO}_{2k+1}(\mathbb{A})$ *to* $\mathrm{GL}_{2k}(\mathbb{A})$ *is injective.*

The point is that, by the strong multiplicity one theorem of Jacquet and Shalika [21] for GL_n, we know that both σ and σ' lift, following [1], to the same automorphic representation τ on $\mathrm{GL}_{2k}(\mathbb{A})$. Now it follows from [1], p. 26, that all twisted local gamma factors of our two representations are the same at all finite places and hence, by Theorem 2.6, they are isomorphic at all finite places. Since the local lift is already prescribed at the archimedean places, σ and σ' are isomorphic at all places. From Theorem 2.6 we derive

Theorem 2.8. *Let τ be in the domain of the descent map σ_ψ from* GL_{2k} *to* SO_{2k+1}. *Then $\sigma_\psi(\tau)$ is irreducible.*

This follows from the fact that $\sigma_\psi(\tau)$ is multiplicity free and all its summands are cuspidal, generic and nearly equivalent, and hence by the rigidity theorem they are all isomorphic and we conclude that $\sigma_\psi(\tau)$ is irreducible.

Let us return to Theorem 2.5. Using the local Langlands reciprocity law for $\mathrm{GL}_n(K)$ proved by Harris–Taylor [17] and Henniart [19], and a theorem of Henniart on the compatibility of the exterior square local L and ε-factors for $\mathrm{GL}_n(K)$ with their corresponding local exterior square Artin factors [20], we obtain

Theorem 2.9. *There exists a unique bijection, preserving twisted ε-factors, between the conjugacy classes of $2k$-dimensional admissible, absolutely irreducible, multiplicity free symplectic representations of the Weil group W_K of K and the isomorphism classes of irreducible, supercuspidal, generic representations of* $\mathrm{SO}_{2k+1}(K)$.

In [24] we extend Theorem 2.8 to a full local Langlands reciprocity law, using Muić's description of all generic representations of $\mathrm{SO}_{2k+1}(K)$ [30].

Theorem 2.10. *For each local Langlands parameter φ of* $\mathrm{SO}_{2k+1}(K)$ (*i.e. a conjugacy class of admissible homomorphisms from $W_K \times \mathrm{SL}_2(\mathbb{C})$ to* $\mathrm{Sp}_{2k}(\mathbb{C})$), *there is a unique, up to isomorphism, irreducible representation $\sigma(\varphi)$ of* $\mathrm{SO}_{2k+1}(K)$ *which is the Langlands subquotient of a parabolic induction of the form $\Sigma(\varphi) = \delta(\Sigma_1) \times \cdots \times \delta(\Sigma_f) \rtimes \sigma^{(t)}$, where $\sigma^{(t)}$ is tempered generic* (*on appropriate* $\mathrm{SO}_{2k'+1}(K)$) *and $\delta(\Sigma_i)$ is an essentially square integrable representation of* $\mathrm{GL}_{n_i}(K)$ *associated to an imbalanced segment Σ_i* (*in the sense of* [37]). *The map $\varphi \mapsto \sigma(\varphi)$ preserves local twisted L and ε-factors. Moreover, $\sigma(\varphi)$ is generic if and only if $\Sigma(\varphi)$ is irreducible.*

As a corollary, we get that, for each tempered local Langlands parameter φ for $\mathrm{SO}_{2k+1}(K)$, the representation $\sigma(\varphi)$ is generic; it will eventually be "the generic member of the tempered local L-packet $\prod(\varphi)$". This is the case of SO_{2k+1} of a conjecture of Shahidi [33]. We also get, as another corollary, a conjecture of Gross–Prasad [16] and of Rallis [27] for $\mathrm{SO}_{2k+1}(K)$.

Theorem 2.11. *With notation as above, $\sigma(\varphi)$ is generic if and only if the local adjoint L-function $L(\mathrm{Ad}_{\mathrm{Sp}_{2k}} \circ \varphi, s)$ is regular at $s = 1$.*

Finally, we get the following applications to automorphic representations.

Theorem 2.12. 1. *Let τ be an irreducible, automorphic, cuspidal representation of* $\mathrm{GL}_{2k}(\mathbb{A})$ *such that $L(\tau, \wedge^2, s)$ has a pole at $s = 1$. Then the local components τ_v are symplectic at all places v.*

2. *The weak lift from irreducible, automorphic, cuspidal, generic representations of* $\mathrm{SO}_{2k+1}(\mathbb{A})$ *to automorphic representations of* $\mathrm{GL}_{2k}(\mathbb{A})$ *is compatible with the local Langlands functorial lift at all places.*

3. *L*-functions for orthogonal groups; non-generic representations

In this section I report on a joint work in progress with S. Rallis towards establishing the existence of a weak functorial lift from irreducible, automorphic, cuspidal representations on a split special orthogonal group, $G_m = \mathrm{SO}_m$ (regarded over F) in m variables, to $\mathrm{GL}_{2[\frac{m}{2}]}$.

Let π be an irreducible, automorphic, cuspidal representation of $G_m(\mathbb{A})$. Consider the Fourier coefficients of Gelfand–Graev type of the form φ_π^{ψ, G'_i} as in Section 1.a, where $G'_i = G_{m-2i-1}$ and φ_π varies in the space of π. These generate, upon restriction to $G'_i(\mathbb{A})$, a space of automorphic functions on $G'_i(\mathbb{A})$, which is invariant to right translations by $G'_i(\mathbb{A})$; denote this space by π^{ψ, G'_i}. Note that the last space is non-trivial for $i = 0$, and if it is non-trivial for $i = \left[\frac{m-1}{2}\right]$, then π is generic. Otherwise let i_1 be the largest index such that $\pi^{\psi, G'_{i_1}}$ is non-trivial. We prove that $\pi^{\psi, G'_{i_1}}$ is cuspidal on $G'_{i_1}(\mathbb{A})$. Choose an irreducible, cuspidal summand of $\pi^{\psi, G'_{i_1}}$ and let ρ_0 be its conjugate representation. Then the $G'_{i_1}(\mathbb{A})$-invariant bilinear pairing $b_{\pi,\rho_0}(\varphi_\pi, \varphi_{\rho_0})$ in (1.3) is non-trivial. We may consider the integrals of the form (1.2),

$$\int_{G_{m-1}(F)\backslash G_{m-1}(\mathbb{A})} \varphi_\pi(g) E(g, f_{s_1,\ldots,s_{i_1};\rho_0})\, dg, \tag{3.1}$$

where $E(g, f_{s_1,\ldots,s_{i_1};\rho_0})$ is an Eisenstein series on $G_{m-1}(\mathbb{A})$ corresponding to the parabolic induction from $|\cdot|^{s_1} \otimes \cdots \otimes |\cdot|^{s_{i_1}} \otimes \rho_0$ on the parabolic subgroup P_1 whose Levi part is isomorphic to $\mathrm{GL}_1^{i_1} \times G_{m-2i_1-1}$. (The difference from (1.2) is that the cuspidal representation τ on GL_{i_1} is replaced by the Eisenstein series induced from the Borel subgroup and its character $|\cdot|^{s_1} \otimes \cdots \otimes |\cdot|^{s_{i_1}}$.) The methods of [9] apply here, as well, and (3.1) is non-trivial and meromorphic. Let us choose and fix $s_1, \ldots, s_{i_1}$ purely imaginary such that (3.1) is non-trivial and $\mathrm{Ind}_{P_1(\mathbb{A})}^{G_{m-1}(\mathbb{A})} |\cdot|^{s_1} \otimes \cdots \otimes |\cdot|^{s_{i_1}} \otimes \rho_0$ is in general position. Denote by ρ the automorphic representation of $G_{m-1}(\mathbb{A})$ generated by the Eisenstein series corresponding to this induced representation. Denote

$$b_{\pi,\rho}(\varphi_\pi, \xi_\rho) = \int_{G_{m-1}(F)\backslash G_{m-1}(\mathbb{A})} \varphi_\pi(g)\xi_\rho(g)\, dg, \tag{3.2}$$

where ξ_ρ is in the space of ρ. Let τ be an irreducible, automorphic, cuspidal representation of $\mathrm{GL}_n(\mathbb{A})$. We consider the integrals (1.1) $\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho})$, even when $i_1 > 0$, in which case ρ is not cuspidal. The methods of [9] still apply, and so we have, for $\mathrm{Re}(s) \gg 0$, the following "Eulerian" expression

$$\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho}) = \int_{G_{m-1}(\mathbb{A})\backslash G_m(\mathbb{A})} \int_{U'_{\mathbb{A}}} b_{\pi,\rho}(\pi(g)\varphi_\pi, f_{\tau_s,\rho}(ug))\psi_{U'}(u)\,du dg, \quad (3.3)$$

where U' is a certain F-unipotent subgroup and $\psi_{U'}$ is a certain character of $U'(\mathbb{A})$, trivial on $U'(F)$. We can find data such that

$$\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho}) = \mathcal{L}_{S_\infty}(\varphi_\pi, f_{\tau_s,\rho}) \frac{L^Z(\pi \times \tau, s)}{L^Z\left(\rho \times \tau, s + \frac{1}{2}\right) L^Z(\tau, r, 2s)}, \quad (3.4)$$

where Z is any finite set of places, including those at infinity, outside which π, ρ_0, τ are all unramified; $r = \wedge^2$ (resp. sym^2) if m is odd (resp. m is even). We denote by S_∞ the set of archimedean primes. $\mathcal{L}_{S_\infty}(\varphi_\pi, f_{\tau_s,\rho})$ is the integral, as in the right-hand side of (3.3), where we replace $\mathbb{A}$ by $\mathbb{A}_{S_\infty}$; it can be chosen to be holomorphic and non-zero at any given point s_0. Fix a finite set of finite places S such that π, ρ_0 are unramified outside $S_\infty \cup S$.

Let τ belong to the twisting set $\mathcal{T}(S, \chi)$ of the converse theorem [3], i.e. χ is a character of $F^*\backslash\mathbb{A}^*$ such that at the places ν of S, χ_ν is highly ramified, and τ_ν is χ_ν times an unramified representation.

Let us define now the (standard) local factors $L(\pi_\nu \times \tau_\nu, s)$ and $\gamma(\pi_\nu \times \tau_\nu, s, \psi_\nu)$ at all places. Let S_τ be a finite set of finite places disjoint from S such that τ is unramified outside $S' = S_\infty \cup S \cup S_\tau$. The definition of the local factors is clear outside S' via the unramified parameters. At the places of S_τ the definition is by multiplicativity of the local factors in the first variable π which is unramified, where for a character μ the local factors for $\mu \times \tau$ are the ones for GL_n. At the places of S we define $L(\pi_\nu \times \tau_\nu, s) = 1$, and we define the local gamma factor by multiplicativity in the second variable τ which is induced from the Borel subgroup, and for a character μ we define $\gamma(\pi_\nu \times \mu, s, \psi_\nu)$ by the doubling method [31], [28]. Finally, at S_∞ we define the local factors through the Langlands classification. We have similar definitions of local factors for $\rho \times \tau$. Denote by $\mathcal{T}_0(S, \chi)$ the subset of elements τ of $\mathcal{T}(S, \chi)$ which are unitary, with central character, which is trivial on the positive real numbers and diagonally embedded inside $\mathbb{A}^*_{S_\infty}$. We prove:

Theorem 3.1. *For $\tau \in \mathcal{T}_0(S, \chi)$, $L^{S_\infty}(\pi \times \tau, s)$ is holomorphic for* $\mathrm{Re}(s) \geq \frac{1}{2}$.

Indeed, taking the (highly ramified at S) character χ as in [1], p. 12, we get that τ is not self-dual, and we conclude as in loc. cit. that $\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho})$ is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$. As in (3.4) we obtain that

$$\mathcal{L}_{S'}(\varphi_\pi, f_{\tau_s,\rho}) \frac{L^{S'}(\pi \times \tau, s)}{L^{S'}\left(\rho \times \tau, s + \frac{1}{2}\right) L^{S'}(\tau, r, 2s)}$$

is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$. Here $\mathcal{L}_{S'}$ is as in the right-hand side of (3.3), with $\mathbb{A}$ replaced by $\mathbb{A}_{S'}$. Next we can find data such that

$$\mathcal{L}_{S'}(\varphi_\pi, f_{\tau_s,\rho}) = \mathcal{L}_{S_\infty \cup S}(\varphi_\pi, f_{\tau_s,\rho}) \prod_{\nu \in S_\tau} \frac{L(\pi_\nu \times \tau_\nu, s)}{L\big(\rho_\nu \times \tau_\nu, s + \frac{1}{2}\big) L(\tau_\nu, r, 2s)}. \tag{3.5}$$

Finally we can find data such that $\mathcal{L}_{S_\infty \cup S}(\varphi_\pi, f_{\tau_s,\rho}) = \mathcal{L}_{S_\infty}(\varphi_\pi, f_{\tau_s,\rho})$. Thus, $\frac{L^{S_\infty}(\pi \times \tau, s)}{L^{S_\infty}(\rho \times \tau, s+\frac{1}{2}) L^{S_\infty}(\tau, r, 2s)}$ is holomorphic, for $\mathrm{Re}(s) \geq \frac{1}{2}$. Since $L^{S_\infty}(\tau, r, 2s)$ is holomorphic for $\mathrm{Re}(2s) \geq 1$, we conclude that $\frac{L^{S_\infty}(\pi \times \tau, s)}{L^{S_\infty}(\rho \times \tau, s+\frac{1}{2})}$ is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$. By induction on m, $L^{S_\infty}(\rho \times \tau, s + \frac{1}{2})$ is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$, and hence so is $L^{S_\infty}(\pi \times \tau, s)$. Note that S can be enlarged so that the process above can be repeated for a finite sequence $\pi, \rho_0, \rho_1, \ldots, \rho_l$ of irreducible cuspidal representations of $G_m(\mathbb{A}), G_{m-2i_1-1}(\mathbb{A}), G_{m-2(i_1+i_2)-2}(\mathbb{A}), \ldots$ such that they are all unramified outside $S_\infty \cup S$ and ρ_i appears in the space of Gelfand–Graev coefficients of ρ_{i-1}. The basic case of induction is when π is generic, and then the theorem is known by [1].

Let M be the intertwining operator corresponding to a Weyl element of $H = G_{2n+m-1}$, which flips X_n^+ and X_n^-. Then we have a functional equation

$$\mathcal{L}(\varphi_\pi, f_{\tau_s,\rho}) = \mathcal{L}(\varphi_\pi, M(f_{\tau_s,\rho})),$$

which unfolds to

$$\begin{aligned} &\mathcal{L}_{S_\infty \cup S}(\varphi_\pi, f_{\tau_s,\rho}) \prod_{\nu \in S_\tau} \mathcal{L}_\nu(\varphi_\pi, f_{\tau_s,\rho}) L^{S'}(\pi \times \tau, s) \\ &\qquad = \tilde{\mathcal{L}}_{S_\infty \cup S}(\varphi_\pi, M_{S_\infty \cup S}(f_{\tau_s,\rho})) \prod_{\nu \in S_\tau} \tilde{\mathcal{L}}_\nu(\varphi_\pi, M_\nu(f_{\tau_s,\rho})) \\ &\qquad\quad \cdot \frac{L^{S'}\big(\rho \times \tau, s - \frac{1}{2}\big) L^{S'}(\tau, r, 2s-1)}{L^{S'}\big(\rho \times \hat{\tau}, \frac{3}{2} - s\big) L^{S'}(\hat{\tau}, r, 2-2s)} L^{S'}(\pi \times \hat{\tau}, 1-s), \end{aligned} \tag{3.6}$$

where at the places $\nu \in S_\tau$, $\mathcal{L}_\nu$ are local analogs of the right-hand side of (3.3) and $\tilde{\mathcal{L}}_\nu$ are obtained from these after a slight modification. Here we use the uniqueness theorem of [25]; for unramified representations π_ν, ρ_ν the space $\mathrm{Hom}_{G_{m-1}(F_\nu)}(\pi_\nu, \hat{\rho}_\nu)$ is one-dimensional. Recall that π_ν and ρ_ν are unramified for $\nu \in S_\tau$. Using this result again, we also prove a local functional equation and compute the proportionality factor at the places of S_τ.

Theorem 3.2. *For $\nu \in S_\tau$ we have*

$$\frac{\gamma(\pi_\nu \times \tau_\nu, s, \psi_\nu)}{\gamma\big(\rho_\nu \times \tau_\nu, s - \frac{1}{2}, \psi_\nu\big) \gamma(\tau_\nu, r, 2s-1, \psi_\nu)} \mathcal{L}_\nu(\varphi_\pi, f_{\tau_s,\rho}) = \tilde{\mathcal{L}}_\nu(\varphi_\pi, M_\nu(f_{\tau_s,\rho})). \tag{3.7}$$

Note that all the local gamma factors in the left-hand side of (3.7) are well defined. We can also prove (3.7) for the places of S, but we prove it as an identity for each local pairing $b_\nu \in \mathrm{Hom}_{G_{m-1}(F_\nu)}(\pi_\nu \otimes \rho_\nu, 1)$, and local integrals $\mathcal{L}_\nu$, and $\tilde{\mathcal{L}}_\nu$, defined using this pairing b_ν. The point is that at S the representation τ is induced from the Borel subgroup, and we have a way to factorize the local integrals, "one GL_1-twist at a time", and then relate the local integrals above for $G_m \times \mathrm{GL}_1$ to the local integrals arising in the doubling method. Thus we get

$$\begin{aligned}\prod_{\nu\in S} \frac{\gamma(\pi_\nu \times \tau_\nu, s, \psi_\nu)}{\gamma\big(\rho_\nu \times \tau_\nu, s-\frac{1}{2}, \psi_\nu\big)\gamma(\tau_\nu, r, 2s-1, \psi_\nu)} \mathcal{L}_{S_\infty \cup S}(\varphi_\pi, f_{\tau_s,\rho}) \\ = \mathcal{L}^{\sim}_{S_\infty \cup S}(\varphi_\pi, M_S(f_{\tau_s,\rho})).\end{aligned} \tag{3.8}$$

Here $\mathcal{L}^{\sim}$ refers to the modification $\sim$ taking place only at S. Note again that all local gamma factors which appear in the left-hand side of (3.8) are well defined. The formal steps of the proof of (3.8) carry on for S_∞ as well. They yield, in the left-hand side of (3.8), the product over all of $S_\infty \cup S$, where, in the case of S_∞, the local factors are the corresponding Artin local gamma factors, and in the right-hand side of (3.8) we have to replace M_S by $M_{S_\infty \cup S}$ and $\mathcal{L}^{\sim}_{S_\infty \cup S}$ by $\tilde{\mathcal{L}}_{S_\infty \cup S}$. However, there are fine details which need to be taken care of in order to justify the formal steps of the proof; this has to do with analytic continuation (in general) of the local integrals. Let us assume that (3.8) is valid, with $S_\infty \cup S$ replacing S, as we just explained, so that we can continue and point out what we have at present, and what is still missing towards a proof of existence of a weak functorial lift from SO_m (split) to $\mathrm{GL}_{2\lfloor \frac{m}{2} \rfloor}$. With this assumption, (3.6)–(3.8) imply that

$$\begin{aligned}L(\pi \times \tau, s) = \varepsilon(\pi \times \tau, s) L(\pi \times \hat{\tau}, 1-s) \frac{L(\tau, r, 2s-1)}{\varepsilon(\tau, r, 2s-1) L(\hat{\tau}, r, 2-2s)} \\ \cdot \frac{L\big(\rho \times \tau, s-\frac{1}{2}\big)}{\varepsilon\big(\rho \times \tau, s-\frac{1}{2}\big) L\big(\rho \times \hat{\tau}, \frac{3}{2}-s\big)}.\end{aligned} \tag{3.9}$$

By the functional equation for $L(\tau, r, s)$, the first quotient in the right-hand side of (3.9) is 1. By induction on m, the second quotient is also 1, the basic case being that where ρ is generic (or just take the trivial cases $m = 0, 1$). This will prove

Theorem 3.3. *Let $\tau \in \mathcal{T}(S, \chi)$ (notation as above). Assume that* (3.8) *is valid , with $S_\infty \cup S$ replacing S (as explained above). Then*

$$L(\pi \times \tau, s) = \varepsilon(\pi \times \tau, s) L(\pi \times \hat{\tau}, 1-s). \tag{3.10}$$

As in [1], define an irreducible representation $\Pi = \otimes \Pi_\nu$ of $\mathrm{GL}_{2\lfloor \frac{m}{2} \rfloor}(\mathbb{A})$ as follows. For $\nu \notin S_\infty \cup S$, Π_ν is the unramified representation (of $\mathrm{GL}_{2\lfloor \frac{m}{2} \rfloor}(F_\nu)$) obtained from π_ν by the local unramified functorial lift. Similarly, for $\nu \in S_\infty$, Π_ν is obtained from π_ν via the Langlands classification. For $\nu \in S$ choose any irreducible, generic, self-dual representation Π_ν of $\mathrm{GL}_{2\lfloor \frac{m}{2} \rfloor}(F_\nu)$ which has a trivial central character. In particular, Π has a trivial central character.

Theorem 3.4. *Assume that χ is highly ramified at S (depending on π only). Then, for all places ν,*

$$L(\Pi_\nu \times \tau_\nu, s) = L(\pi_\nu \times \tau_\nu, s), \tag{3.11}$$

$$\gamma(\Pi_\nu \times \tau_\nu, s, \psi_\nu) = \gamma(\pi_\nu \times \tau_\nu, s, \psi_\nu), \tag{3.12}$$

Similar identities apply to $\hat{\tau}$.

By construction, (3.11) and (3.12) are clear for all ν outside S. For $\nu \in S$, since χ_ν is highly ramified, both sides of (3.11) are 1. As for (3.12), let us write τ_ν as the representation, induced from the Borel subgroup and a character $\mu_{1,\nu}\chi_\nu \otimes \cdots \otimes \mu_{n,\nu}\chi_\nu$, where $\mu_{1,\nu}, \ldots, \mu_{n,\nu}$ are unramified. Then we need to prove

$$\prod_{i=1}^{n} \gamma(\Pi_\nu \times \mu_{i,\nu}\chi_\nu, s, \psi_\nu) = \prod_{i=1}^{n} \gamma(\pi_\nu \times \mu_{i,\nu}\chi_\nu, s, \psi_\nu).$$

For this it is enough to prove

$$\gamma(\Pi_\nu \times \chi_\nu, s, \psi_\nu) = \gamma(\pi_\nu \times \chi_\nu, s, \psi_\nu). \tag{3.13}$$

Recall again that the right-hand side of (3.13) is defined via the doubling method. Both sides of (3.13) are stable for sufficiently ramified χ_ν. See [22] for the stability of the left-hand side, and [32] for the stability of the right-hand side. Thus we may replace π_ν and Π_ν with a pair of unramified representations, which correspond without changing their local gamma factors (twisted by χ_ν). This proves (3.13).

The main property which is missing at this stage is the holomorphicity of the full L-function $L(\pi \times \tau, s)$ for $\mathrm{Re}(s) \geq \frac{1}{2}$. Recall from Theorem 3.1 that we know that $L^{S_\infty}(\pi \times \tau, s)$ is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$, when $\tau \in \mathcal{T}_0(S, \chi)$. Assume, for example, that π is tempered at S_∞. Then $\prod_{\nu \in S_\infty} L(\pi_\nu \times \tau_\nu, s)$ is holomorphic for $\mathrm{Re}(s) \geq \frac{1}{2}$, and hence so is $L(\pi \times \tau, s)$. In such a case Theorem 3.3 (where we assumed that (3.8) is valid for $S_\infty \cup S$ as well) will imply that $L(\pi \times \tau, s)$ is entire. Once we have this, we see that by Theorem 2 in [5], $L(\pi \times \tau, s)$ is an entire function of finite order, and we can conclude that it is bounded in vertical strips (and similarly for $\hat{\tau}$). Then we can apply the converse theorem and obtain an automorphic representation of $\mathrm{GL}_{2[\frac{m}{2}]}(\mathbb{A})$ which is isomorphic to Π at all places outside S. Finally, let us mention that the ideas of the descent method can be applied here as well, and moreover, cuspidal representations on any orthogonal group (split or otherwise) can be considered along similar lines. These topics are the subject of current work in progress, which we hope to report on in future times.

References

[1] Cogdell, J., Kim, H., Piaretski-Shapiro, I., Shahidi, F., On lifting from classical groups to GL_N. *Inst. Hautes Études Sci. Publ. Math.* **93** (2001), 5–30.

[2] Cogdell, J., Kim, H., Piaretski-Shapiro, I., Shahidi, F., Functoriality for classical groups. *Inst. Hautes Études Sci. Publ. Math.* **99** (2004), 163–233.

[3] Cogdell, J., Piatetski-Shapiro, I., Converse theorems for GL_n. *Inst. Hautes Études Sci. Publ. Math.* **79** (1994), 157–214.

[4] Friedberg, S., Goldberg, D., On local coefficients for non-generic representations of some classical groups. *Compositio Math.* **116** (1999), 133–166.

[5] Gelbart, S., Lapid, E., Lower bounds for L-functions at the edge of the critical strip. Preprint, 2005.

[6] Gelbart, S., Piatetski-Shapiro, I., L-functions for $G \times \mathrm{GL}_n$. In *Explicit Constructions of Automorphic L-functions*, Lecture Notes in Math. 1254, Springer-Verlag, Berlin 1987, 53–143.

[7] Gelbart, S., Shahidi, F., *Analytic properties of automorphic L-functions*. Perspectives in Mathematics 6, Academic Press Inc., Boston, MA, 1988.

[8] Ginzburg, D., L-functions for $\mathrm{SO}_n \times \mathrm{GL}_k$. *J. Reine Angew. Math.* **405** (1990), 156–180.

[9] Ginzburg, D., Piatetski-Shapiro, I., Rallis, S., *L-functions for the orthogonal group*. Mem. Amer. Math. Soc. 128, no. 611, Amer. Math. Soc., Providence, RI, 1997.

[10] Ginzburg, D., Rallis, S., Soudry, D., On explicit lifts of cusp forms from $\mathrm{GL}(m)$ to classical groups. *Ann. of Math.* **150** (1999), 807–866.

[11] Ginzburg, D., Rallis, S., Soudry, D., On a correspondence between cuspidal representations of GL_{2n} and $\tilde{\mathrm{S}}\mathrm{p}_{2n}$. *J. Amer. Math. Soc.* **12** (1999), 849–907.

[12] Ginzburg, D., Rallis, S., Soudry, D., Endoscopic representations of $\tilde{\mathrm{S}}\mathrm{p}_{2N}$. *J. Inst. Math. Jussieu* **1** (2002), 77–123.

[13] Ginzburg, D., Rallis, S., Soudry, D., Lifting cusp forms on GL_{2n} to $\tilde{\mathrm{S}}\mathrm{p}_{2n}$: the unramified correspondence. *Duke Math. J.* **100**(2) (1999), 243–266.

[14] Ginzburg, D., Rallis, S., Soudry, D., Generic automorphic forms on $\mathrm{SO}(2n+1)$: functorial lift to $\mathrm{GL}(2n)$, endoscopy and base change. *Internat. Math. Res. Notices* **14** (2001), 729–764.

[15] Ginzburg, D., Rallis, S., Soudry, D., L-functions for symplectic groups. *Bull. Soc. Math. France* **126** (1998), 181–244.

[16] Gross, B., Prasad, D., On the decomposition of a representation of SO_n when restricted to SO_{n-1}. *Canad. J. Math.* **44** (1992), 974–1002.

[17] Harris, M., Taylor, R., *On the geometry and cohomology of some simple Shimura varieties*. Annals of Mathematical Studies 151, Princeton University Press, Princeton, NJ, 2001.

[18] Henniart, G., Caractérization de la correspondance de Langlands locale par les facteurs ε de paires. *Invent. Math.* **113** (1993), 339–350.

[19] Henniart, G., Une preuve simple des conjectures de Langlands pour $\mathrm{GL}(n)$ sur un corps p-adique. *Invent. Math.* **139** (2000), 439–455.

[20] Henniart, G., Correspondance de Langlands et fonctions L des carrés extérieure et symmétrique. Preprint, Inst. Hautes Études Sci., 2003.

[21] Jacquet, H., Shalika, J., On Euler products and the classification of automorphic forms I, II. *Amer. J. Math.* **103** (1981), 499–558, 777–815.

[22] Jacquet, H., Shalika, J., A lemma on highly ramified ε-factors. *Math. Ann.* **271**, (1985), 319–332.

[23] Jiang, D., Soudry, D., The local converse theorem for $\mathrm{SO}(2n+1)$ and applications. *Ann. of Math.* **157** (2003), 743–806.

[24] Jiang, D., Soudry, D., Generic representations and local Langlands reciprocity law for p-adic SO_{2n+1}. In *Contributions to Automorphic Forms, Geometry and Number Theory, A Volume in Honor of Joseph A. Shalika* (ed. by H. Hida, D. Ramakrishnan and F. Shahidi), The Johns Hopkins University Press, Baltimore, MD, 2004, 457–519.

[25] Kato, S.-I., Murase, A., Sugano, T., Whittaker-Shintani functions for orthogonal groups. *Tohoku Math. J.* **55**, no. 1 (2003), 1–64.

[26] Kim, H., Krishnamurthy, M., Stable base change lift from unitary groups to GL_N. *IMRP Int. Math. Res. Pap.* **1** (2005), 1–52.

[27] Kudla, S., The local Langlands correspondence: the non-archimedean case. In *Motives* (ed. by U. Jannsen, S. Kleiman, J.-P. Serre). Proc. Symp. Pure Math. 55, part 2, Amer. Math. Soc., Providence, RI, 1994, 365–397.

[28] Lapid, E., Rallis, S., On the local factors for classical groups. In *Automorphic Representations, L-Functions and Applications: Progress and Prospects* (ed. by J. W. Cogdell, D. Jiang, S. S. Kudla, D. Soudry, R. Stanton), Proceedings of a conference in honor of Steve Rallis, Ohio State Univ. Math. Res. Inst. Publ. 11, Walter de Gruyter, Berlin 2005, 309–360.

[29] Moeglin, C., Waldspurger, J.-L., Modèles de Whittaker dégénérés pour des groupes p-adiques. *Math. Z.* **196** (1987), 427–452.

[30] Muić, G., On generic irreducible representations of $\mathrm{Sp}(n, F)$, $\mathrm{SO}(2n+1, F)$. *Glas. Mat. Ser. III* **33** (53) (1998), 19–31.

[31] Piatetski-Shapiro, I., Rallis, S., L-functions for the classical groups. In *Explicit Constructions of Automorphic L-functions*, Lecture Notes in Math. 1254, Springer-Verlag, Berlin 1987, 1–52.

[32] Rallis, S., Soudry, D., Stability of the local gamma factor arising from the doubling method. *Math. Ann.* **333** (2005), 291–313.

[33] Shahidi, F., A proof of Langlands conjecture on Plancherel measures; Complementary series for p-adic groups. *Ann. of Math.* **132** (1990), 273–330.

[34] Soudry, D., On Langlands functoriality from classical groups to GL_n. In *Formes Automorphes* (I)*; Actes du Semestre du CEB, Printemps* 2000 (ed. by J. Tilouine, H. Carayol, M. Harris, M.-F. Vignéras), Astérisque **298** (2005), 335–390.

[35] Tamir, B., On L-functions and intertwining operators for unitary groups. *Israel J. Math.* **73** (1991), 161–188.

[36] Watanabe, T., A comparison of automorphic L-functions in a theta series lifting for unitary groups. *Israel J. Math.* **116** (2000), 93–116.

[37] Zelevinsky, A.V., Induced representations of reductive p-adic groups II. On irreducible representations of $\mathrm{GL}(n)$. *Ann. Sci. École Norm. Sup.* **13** (1980), 165–210.

School of Mathematical Sciences, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

E-mail: soudry@math.tau.ac.il

Representation theory and the cohomology of arithmetic groups

Birgit Speh*

Abstract. Let G be a semisimple Lie group with finitely many connected components and Lie algebra $\mathfrak{g}$, K a maximal compact subgroup of G, and $X = G/K$ a symmetric space. A torsion free discrete subgroup Γ of G and a finite dimensional real or complex linear representation (ρ, E) of G define a locally symmetric space $X_\Gamma = \Gamma\backslash G/K$ with a local system $\tilde{E}$. Then $H^*(\Gamma, E) = H^*(\Gamma\backslash X, \tilde{E})$ is isomorphic to $H^*(\mathfrak{g}, K, C^\infty(\Gamma\backslash G) \otimes E)$. If Γ is an arithmetic group, then $H^*(\mathfrak{g}, K, C^\infty(\Gamma\backslash G) \otimes E)$ is isomorphic to the $(\mathfrak{g}, K)$-cohomology with coefficients in $\mathcal{A}(\Gamma\backslash G) \otimes E$ where $\mathcal{A}(\Gamma\backslash G)$ is the space of automorphic forms. Using representation theory and the theory of automorphic forms a large amount of information about $H^*(\Gamma, E)$ can be deduced.

Mathematics Subject Classification (2000). Primary 22E40; Secondary 22E46.

Keywords. Cohomology, arithmetic groups, unitary representations, reductive Lie groups.

1. Introduction

Let $\boldsymbol{G}$ be a semi-simple simply connected algebraic group over $\mathbb{Q}$, K be a maximal compact subgroup of its real points $G = \boldsymbol{G}(\mathbb{R})$ and $X = G/K$ the global symmetric space. Assume that $\Gamma \subset \boldsymbol{G}(\mathbb{Q})$ is a torsion-free congruence subgroup. Then $S(\Gamma) = \Gamma\backslash X$ is a manifold with finite volume under the G-invariant metric and volume form on $X = G/K$.

Suppose that $\boldsymbol{H}$ is a $\mathbb{Q}$-rational reductive subgroup with real points H such that $K \cap H = K_H$ is maximal compact in H. Then the inclusion

$$X_H := H/K_H \to X$$

induces a map

$$j\colon \Gamma \cap H\backslash X_H \longrightarrow \Gamma\backslash X.$$

We assume from now on that $\Gamma\backslash X$ and $S_H(\Gamma) := \Gamma \cap H\backslash X_H$ are not compact.

Let $[\phi]$ be an element in the i-th cohomology with compact support $H^i_c(S(\Gamma), \mathbb{C})$ represented by a closed compactly supported i-form ϕ and that $i + k =$

*The author is supported by NSF grant A78-8332-6820.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$\dim(X_H)$. Then $[\phi]$ determines a map

$$(\phi, S_H)\colon H^k(S(\Gamma), \mathbb{C}) \to \mathbb{C},$$
$$[\omega] \mapsto \int_{\Gamma\cap H\backslash X_H} \phi \wedge j^*\omega.$$

We call this map the modular symbol attached to $([\phi], S_H)$. If we can find $[\omega] \in H^*(S(\Gamma), \mathbb{C})$ such that $([\phi], H)([\omega]) \neq 0$ then $([\phi], S_H)$ is a nontrivial modular symbol. If $S(\Gamma)$ is oriented we use Poincaré duality and identify $([\phi], S_H)$ with an element in $H_c^*(S(\Gamma), \mathbb{C})$.

On the other hand if $[\psi] \in H^i(\Gamma \cap H\backslash X_H, \mathbb{C})$ and $i + k = \dim(X_H)$ the map

$$([\psi], S_H)\colon H_c^k(\Gamma\backslash X, \mathbb{C}) \to \mathbb{C},$$
$$[\omega] \mapsto \int_{\Gamma\cap H\backslash X_H} \psi \wedge j^*\omega$$

can be identified with an element in $H^*(\Gamma\backslash X, \mathbb{C})$ which we also call a modular symbol attached to $([\psi], S_H)$. If the degree of ψ is 0 then we say that the fundamental class $[S_H(\Gamma)]$ is a modular symbol in $H^{\dim X - \dim X_H}(S(\Gamma), \mathbb{C})$.

The Hecke algebra acts on $H^*(S(\Gamma), \mathbb{C})$ and on $H_c^*(S(\Gamma), \mathbb{C})$ and one may consider the span of Hecke translates of the modular symbols $([\psi], S_H)$ and in particular $[S_H(\Gamma)]$, where H ranges over all groups H. It is unknown, how much of the cohomology of $S(\Gamma)$ may be captured by nonzero modular symbols and their Hecke translates.

The cohomology $H^*(\Gamma\backslash X, \mathbb{C})$ of the locally symmetric space $\Gamma\backslash X$ can be identified with the $(\mathfrak{g}, K)$-cohomology of the G-module $\mathcal{A}(\Gamma\backslash G)$ of automorphic functions on $\Gamma\backslash G$. Thus it is determined by representation theoretic and arithmetic data. Understanding the relationship between nontrivial modular symbols attached to a subgroup $\boldsymbol{H}$ and the subrepresentations of $\mathcal{A}(\Gamma\backslash G)$ gives important arithmetic information, since many of the integrals $\int_{\Gamma\cap H\backslash X_H} \psi \wedge j^*\omega$ are period integrals or are related to special values of L-functions. This is discussed in [7], [11], [12], [4] and the recent work of S. Kudla [8].

For some compact locally symmetric spaces Kudla and Millson [9], by Clozel and Venkataranama [4] and by Tong and Wang [19], have obtained results relating some special $[S_H(\Gamma)]$ to cohomology classes defined by automorphic representations. I will discuss here some connections between cohomology classes defined using automorphic representations and nontrivial modular symbols for noncompact locally symmetric spaces.

2. The cohomology $H^*(S(\Gamma), \mathbb{C})$ and automorphic representations

We denote the Lie algebra by gothic letters and write $\mathfrak{g} = \mathfrak{k} + p$ for the Cartan decomposition of the Lie algebra $\mathfrak{g}$. The complex of the differential forms on $S(\Gamma)$ is

isomorphic to

$$\mathrm{Hom}_K(\wedge^*\mathfrak{g}/\mathfrak{k}, C^\infty(\Gamma\backslash G))$$

(see [3]). Here consider $C^\infty(\Gamma\backslash G)$ as a left $\mathfrak{g}$-module and

$$df(x_0,\dots x_q) = \sum_i x_i f(x_0,\dots,\hat{x}_i,\dots x_q) + \sum_{i<j}(-1)^{i+j} f([x_i,x_j],x_0,\dots,\hat{x}_i,\dots,\hat{x}_j,\dots,x_q).$$

We denote its cohomology by $H^*(S(\Gamma),\mathbb{C}) = H^*(\mathfrak{g},K,C^\infty(\Gamma\backslash G))$. Similarly we define for an admissible representation π of G the $(\mathfrak{g},K)$-cohomology $H^*(\mathfrak{g},K,\pi)$.

We denote the space of automorphic functions on $\Gamma\backslash G$ by $\mathcal{A}(\Gamma\backslash G)$. An admissible $(\mathfrak{g},K)$-module, which is isomorphic to a subrepresentation of $\mathcal{A}(\Gamma\backslash G)$, is called an automorphic representation. The space of square integrable automorphic functions is a direct sum of $\mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)$, the space of residual automorphic functions and $\mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G)$, the space of the cuspidal automorphic functions [10]. An irreducible unitary $(\mathfrak{g},K)$-submodule of $\mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)$, respectively of $\mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G)$ is called a residual, respectively cuspidal $(\mathfrak{g},K)$-module.

As a $(\mathfrak{g},K)$-module

$$\mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G) = \oplus_{U\in\hat{G}_u} m(\Gamma,U)U.$$

Here $\hat{G}_u$ denotes the set of all unitary irreducible $(\mathfrak{g},K)$-modules and

$$m(\Gamma,U) = \dim \mathrm{Hom}_{\mathfrak{g},K}(U,\mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G)).$$

A result by A. Borel [2] states that if π is a cuspidal $(\mathfrak{g},K)$-module then

$$H^*(\mathfrak{g},K,\pi) \hookrightarrow H^*(\mathfrak{g},K,C^\infty(\Gamma\backslash G)) = H^*(S(\Gamma),\mathbb{C}). \tag{1}$$

In particular, if $S(\Gamma)$ is compact, then

$$\mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G) = \mathcal{A}(\Gamma\backslash G)$$

and hence

$$H^*(S(\Gamma),\mathbb{C}) = H_c^*(S(\Gamma),\mathbb{C}) = \oplus_{U\in\hat{G}_u} m(\Gamma,U)H^*(\mathfrak{g},K,U).$$

See [3] for details.

If $\Gamma\backslash G$ and hence $S(\Gamma)$ are not compact a result of Franke [5] shows that

$$H^*(S(\Gamma),\mathbb{C}) = H^*(\mathfrak{g},K,\mathcal{A}(\Gamma\backslash G)).$$

In contrast to (1)

$$H^*(\mathfrak{g},K,\mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)) \not\hookrightarrow H^*(\mathfrak{g},K,C^\infty(\Gamma\backslash G))$$

as the following example illustrates.

The trivial representation I of G is an automorphic representation satisfying $H^*(\mathfrak{g}, K, I) \neq 0$. Let G_u be the compact subgroup of $\boldsymbol{G}(\mathbb{C})$ with Lie algebra $\mathfrak{k} \oplus ip$ and write $\hat{X} = G_u/K$ for the compact dual of X. We identify $H^*(\hat{X})$ with $H^*(\mathfrak{g}, K, I)$. The Borel map j from the cohomology $H^*(\hat{X})$ into the cohomology $H^*(S(\Gamma), \mathbb{C})$ is defined by the inclusion of the constant functions $\mathbb{C}$ in the space $\mathcal{A}(\Gamma\backslash G)$. The classes in the image of $H^*(\mathfrak{g}, K, I)$ in $H^*(S(\Gamma), \mathbb{C})$ are represented by invariant differential forms. If D is the dimension of $S(\Gamma)$ then $\dim H^D(\mathfrak{g}, K, I) = 1$. On the other hand $H^D(S(\Gamma), \mathbb{C}) = 0$ since $S(\Gamma)$ is not compact.

The kernel of the Borel map $j\colon H^*(\mathfrak{g}, K, I) \to H^*(S(\Gamma), \mathbb{C})$ is determined in [6].

According to Parthasaraty, Kumaresan and Vogan–Zuckerman the irreducible unitary representations π with $H^*(\mathfrak{g}, K, \pi) \neq 0$ are classified as follows: let $\mathfrak{h} = \mathfrak{t} \oplus \mathfrak{a}$ be a θ-stable fundamental Cartan subalgebra and $\mathfrak{q}_{\mathbb{C}} = \mathfrak{l}_{\mathbb{C}} \oplus \mathfrak{n}_{\mathbb{C}}$ a θ-stable parabolic subgroup of $\mathfrak{g} \otimes \mathbb{C}$ containing $\mathfrak{h} \otimes \mathbb{C}$. Then $\mathfrak{l}_{\mathbb{C}}$, $\mathfrak{l}$ and L are the centralizers of an element $X \in i\mathfrak{t}$ in $\mathfrak{g}_{\mathbb{C}}$, $\mathfrak{g}$ and G respectively. Associated to $\mathfrak{q}$ is an irreducible unitary $(\mathfrak{g}, K)$-modules $A_{\mathfrak{q}}$ and an irreducible finite dimensional representation $V(\mathfrak{q}) \subset \wedge^{\dim(p \cap \mathfrak{n})} p$ of K so that

$$H^{\dim(p \cap \mathfrak{n})}(\mathfrak{g}, KA_{\mathfrak{q}}) \cong \mathrm{Hom}(V(\mathfrak{q}), A_{\mathfrak{q}}) \cong \mathbb{C},$$

and we obtain forms representing all other nontrivial cohomology classes as the wedge product of $0 \neq \omega_{\mathfrak{q}} \in \mathrm{Hom}(V(\mathfrak{q}), A_{\mathfrak{q}})$ with forms representing classes in $H^*(\mathfrak{l}, L \cap K, I)$. Every irreducible unitary representation with nontrivial $(\mathfrak{g}, K)$-cohomology is isomorphic to a representation $A_{\mathfrak{q}}$.

Suppose that $\boldsymbol{P}$ is a rational parabolic subgroup and that $A_{\mathfrak{q}}$ is the Langlands quotient of $I(\boldsymbol{P}(\mathbb{R}), \pi, \nu)$. Assume furthermore that there is a map

$$E_{\mathrm{res}}\colon I(\boldsymbol{P}(\mathbb{R}), \pi, \nu) \to \mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)$$

which defines a $(\mathfrak{g}, K)$-map $J\colon A_{\mathfrak{q}} \hookrightarrow \mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)$ into the residual spectrum. It extends to a map

$$J^*\colon \mathrm{Hom}_K(\wedge^* p, A_{\mathfrak{q}}) \to \mathrm{Hom}_K(\wedge^* p, \mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)).$$

One can show that $[J^*(\omega_{\mathfrak{q}}] \neq 0$.

I expect the following generalization of Franke's result to hold.

Conjecture. Let $[\omega_L] \in H^*(\mathfrak{l}, K \cap L, I)$ be in the kernel of the Borel map for L. Then

$$[J^*(\omega_{\mathfrak{q}} \wedge \omega_L)] = 0.$$

The contribution to $H^*(S(\Gamma), \mathbb{C})$ by representations with nontrivial $(\mathfrak{g}, K)$-cohomology which are induced from parabolic subgroups and which are embedded via Eisenstein series into $\mathcal{A}(\Gamma\backslash G)$ has been considered by G. Harder and J. Schwermer [15].

3. Locally symmetric spaces in the adelic language

Let $\boldsymbol{A}$ be the adeles of $\mathbb{Q}$ and $\boldsymbol{A}_f$ the finite adeles. For a linear algebraic semi-simple simply connected group $\boldsymbol{G}$ defined over $\mathbb{Q}$ we define $\boldsymbol{G}(\boldsymbol{A})$, $\boldsymbol{G}(\mathbb{Q})$, $\boldsymbol{G}(\boldsymbol{A}_f)$ and a compact group

$$K_{\boldsymbol{A}} = K_\infty \prod_p \boldsymbol{G}(\mathcal{O}_p) = K_\infty K_{\boldsymbol{A}_f},$$

where K_∞ is the maximal compact subgroup of $G = \boldsymbol{G}(\mathbb{R})$. We have a locally symmetric space

$$S(K_{\boldsymbol{A}_f}) = \boldsymbol{G}(\mathbb{Q})\backslash X \times \boldsymbol{G}(\boldsymbol{A}_f)/K_{\boldsymbol{A}_f}.$$

For each subgroup $K'_{\boldsymbol{A}_f} \subset K_{\boldsymbol{A}_f}$ of finite index we also define a locally symmetric space $S(K'_{\boldsymbol{A}}) = \boldsymbol{G}(\mathbb{Q})\backslash X \times \boldsymbol{G}(\boldsymbol{A}_f)/K'_{\boldsymbol{A}_f}$. If K'_f is sufficiently small then

$$S(K'_f) = \bigcup_{i=1}^{h} \Gamma_i \backslash X$$

is a finite disjoint union of locally symmetric spaces $\Gamma_i\backslash X$, where the $\Gamma_i \subset G(\mathbb{Q})$ are congruence subgroups. Furthermore

$$H^*(S(K'_f), \mathbb{C}) \cong H^*(\mathfrak{g}, K_\infty, \mathcal{A}(\boldsymbol{G}(\mathbb{Q})\backslash G(\boldsymbol{A}))^{K'_f}.$$

We give $\boldsymbol{G}(\boldsymbol{A}_f)$ the topology induced by the topology of $\boldsymbol{A}_f$ and define $X_{\boldsymbol{A}} := X \times \boldsymbol{G}(\boldsymbol{A}_f)$. Then $X_{\boldsymbol{A}}$ with the product topology is the *adelic symmetric space* attached to $\boldsymbol{G}$. The group $\boldsymbol{G}(\mathbb{Q})$ of $\mathbb{Q}$-rational points of $\boldsymbol{G}$ acts by left translation freely and discontinuously on $X_{\boldsymbol{A}}$. The space $\boldsymbol{G}(\mathbb{Q})\backslash X_{\boldsymbol{A}} =: S_G$ with the quotient topology is called the *adelic locally symmetric space* attached to $\boldsymbol{G}$.

Let $\tilde{\mathbb{C}}$ be a locally constant sheaf S_G. By $H^*(S_G, \tilde{\mathbb{C}})$ we denote the smooth sheaf cohomology of S_G with coefficients $\tilde{\mathbb{C}}$. The group $\boldsymbol{G}(\boldsymbol{A}_f)$ acts by right translation on S_G and $H^j(S_G, \tilde{\mathbb{C}})$ is a smooth $\boldsymbol{G}(\boldsymbol{A}_f)$-module, i.e. if K_f runs in the set of compact open subgroups of $\boldsymbol{G}(\boldsymbol{A}_f)$ and if $H^j(S_G, \tilde{\mathbb{C}})^{K_f}$ denotes the K_f-invariants in $H^j(S_G, \tilde{\mathbb{C}})$ then

$$\bigcup_{K_f} H^j(S_G, \tilde{\mathbb{C}})^{K_f} = H^j(S_G, \tilde{\mathbb{C}}).$$

Moreover

$$H^j(S_G, \tilde{\mathbb{C}})^{K_f} = H^j(S_G/K_f, \tilde{\mathbb{C}})$$

where $S_G/K_f = S(K_f)$ is the topological quotient of S_G by the K_f-action, and $H^j(S_G/K_f, \tilde{\mathbb{C}})$ is the sheaf-cohomology of S_G/K_f with coefficients in the sheaf $\tilde{\mathbb{C}}$.

A finite dimensional representation $\rho\colon G \to \mathrm{End}(V)$ of G defines a locally constant sheaf $\tilde{V}$ on S_G and we have similar formulas for the sheaf $\tilde{V}$.

4. Modular symbols and automorphic representations

Consider first the trivial representation I which is in the residual spectrum of $\mathcal{A}(\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A}))$.

For a subgroup $\boldsymbol{H}$ define the compact dual $\hat{X}_H$ of X_H analogously. One has an embedding of $\hat{X}_H$ in $\hat{X}$, and the fundamental class of $[\hat{X}_H]$ is a cohomology class in $H^*(\hat{X})$.

Put $\Gamma_0 = \boldsymbol{G}(\mathbb{Q}) \cap \prod_p \boldsymbol{G}(\mathcal{O}_p)$ and let $\mathcal{H}_0$ to be the space of complex valued compactly supported $K_f = \prod_p \boldsymbol{G}(\mathcal{O}_p)$-bi-invariant functions on the finite adelic group $G(\boldsymbol{A}_f)$. This is the Hecke algebra corresponding to K_f under convolutions and acts on the cohomology group $H^*(S(\Gamma_0), \mathbb{C}) = H^*(S_G, \tilde{\mathbb{C}})^{K_f}$.

Theorem ([17]). *Suppose that $\boldsymbol{G}$ is a simply connected group which has no $\mathbb{R}$-anisotropic factors defined over $\mathbb{Q}$. Then, the class $j([\hat{X}_H])$ is a linear combination of Hecke translates of the generalized modular symbol $[S_H(\Gamma')] \in H^*(S(\Gamma'), \mathbb{C})$ for some congruence subgroup Γ' of Γ_0.*

In particular, if $j([\hat{X}_H]) \neq 0$, then the modular symbol $[S_H(\Gamma')]$ does not vanish.

Since the kernel of the Borel map j has been determined in [6], this result is used to exhibit many modular symbols in [17] related to the trivial representation.

We now consider representations with nontrivial $(\mathfrak{g}, K)$-cohomology which are induced from parabolic subgroups and which are considered as subrepresentations of $\mathcal{A}(\boldsymbol{G}(\mathbb{Q})\backslash\boldsymbol{G}(\boldsymbol{A}))$ via Eisenstein series. Let $\boldsymbol{P} \supset \boldsymbol{B}$ be a standard parabolic subgroup of $\boldsymbol{G}$ with standard Levi part $\boldsymbol{L}_P$. Let $x_0 = K\,e \in X$. The orbit of x_0 under $L_P(\boldsymbol{A})$ in the globally symmetric space $X_{\boldsymbol{A}}$ is isomorphic to $\boldsymbol{L}_P(\mathbb{R})/(\boldsymbol{L}_P(\mathbb{R}) \cap K)) \times \boldsymbol{L}_P(\boldsymbol{A}_f)$. We have a symmetric space $\boldsymbol{L}_P(\mathbb{R})/(\boldsymbol{L}_P(\mathbb{R}) \cap K)$ and

$$S^{\natural}_{L_P} := \boldsymbol{L}_P(\mathbb{Q})\backslash\boldsymbol{L}_P(\mathbb{R})/(\boldsymbol{L}_P(\mathbb{R}) \cap K)) \times \boldsymbol{L}_P(\boldsymbol{A}_f)$$

is a locally symmetric space and we have a continuous injection $S^{\natural}_{L_P} \hookrightarrow S_G$, which identifies it with a closed subspace of S_G. We call $S^{\natural}_{L_P}$ the *modular manifold* attached to P.

In [14] we consider the covering space $S_P := \boldsymbol{P}(\mathbb{Q})\backslash X_{\boldsymbol{A}}$. One can see that

$$H^*(S_P, \mathbb{C}) = H^*(\boldsymbol{P}(\mathbb{Q}), C^\infty(\boldsymbol{G}(\boldsymbol{A}_f), \mathbb{C})).$$

Using the Steinberg representation we define the Poincaré dual $H^*_c(S_P, \mathbb{C})$. In [13] we attach to a class $[\phi] \in H^j_c(S_P, \mathbb{C})$ a class

$$\mathrm{cor}_P([\phi]) \in H^j_c(S_G, \tilde{V}).$$

The classes are in a sense the Poincaré dual to Eisenstein classes defined by embeddings of representations induced from $\boldsymbol{P}(\boldsymbol{A})$ into the space of automorphic functions [15]. In [14] we obtain formulas for a restriction map res_c of the classes $\mathrm{cor}_P([\phi])$

to $H^*(S^{\natural}_{L_P}, \mathbb{C})$, thus obtaining modular symbols. As a special case we prove that the fundamental class $[S^{\natural}_{L_P}]$ defines a nontrivial modular symbol generalizing a result by Ash and Borel [1].

For $G = \mathrm{Gl}(2n, \mathbb{R})$ there is a unique unitary irreducible representation $J(n)$ with

$$H^{n(n+1)/2}(\mathfrak{g}, K, J(n)) \neq 0$$

and

$$H^j(\mathfrak{g}, K, J(n)) = 0 \quad \text{for } j < n(n+1)/2.$$

By Poincaré duality $H^{(n+1)(3n-2)/2}(\mathfrak{g}, K, J(n)) \neq 0$.

If Γ is a small congruence subgroup the residues of Eisenstein series define an embedding

$$E\colon J(n) \to \mathcal{A}_{\mathrm{res}}(\Gamma\backslash G)$$

in the residual spectrum and a nonzero class in $H^{n(n+1)/2}(S(\Gamma), \mathbb{C})$ [16]. By Poincaré duality we have a nonzero class in $H_c^{(n+1)(3n-2)/2}(S(\Gamma), \mathbb{C})$ which can be represented by a pseudo Eisenstein form ω_{sp} [13].

Now let $H = \mathrm{Sp}(n, \mathbb{R}) \subset \mathrm{Gl}(2n, \mathbb{R})$ be defined by the skew symmetric form $\left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. The work of Jacquet and Rallis implies that $J(n)$ admits an H-invariant linear functional and no other irreducible unitary infinite dimensional representation with nontrivial $(\mathfrak{g}, K)$-cohomology admits an H-invariant linear functional.

Proposition. *Suppose that $n = 2$. If Γ is a sufficiently small congruence subgroup, then $\int_{S_H(\Gamma)} \omega_{\mathrm{sp}} \neq 0$ and so the modular symbol $[S_H(\Gamma)]$ is not zero.*

Suppose that $n > 2$. Then the Oda restriction of $[\omega_{\mathrm{sp}}]$ to $S(\Gamma)$ is zero.

5. A conjecture

These examples support the conjecture that understanding the automorphic representations with nontrivial $(\mathfrak{g}, K)$-cohomology as H- modules and their $(\mathfrak{h}, K_H)$-quotients, respectively subrepresentations, is crucial to understanding their relationship to modular symbols. In general the restriction of π to $\mathfrak{h}$ is not a direct sum of unitary irreducible $(\mathfrak{h}, K_H)$-modules and so we have to consider its irreducible quotients.

Suppose that $j\colon H \hookrightarrow G$ is a subgroup, π an irreducible unitary automorphic representation with $H^*(\mathfrak{g}, K, \pi) \neq 0$. Let $0 \neq [\omega] \in H^*(\mathfrak{g}, K, \pi)$. Assume also that there exists an irreducible $(\mathfrak{h}, K_H)$-module π_H and $Q \in \mathrm{Hom}_H(\pi, \pi_H)$ and $0 \neq [Q^*\omega] \in H^*(\mathfrak{h}, K \cap H, \pi_H)$.

Let

$$E\colon \pi \hookrightarrow \mathcal{A}_{\mathrm{res}}(\Gamma\backslash G) \oplus \mathcal{A}_{\mathrm{cusp}}(\Gamma\backslash G)$$

be an embedding.

Conjecture. Suppose that $[E^*(\omega)] \neq 0$. For sufficiently small congruence subgroup Γ' the Oda restriction of $[E^*\omega]$ to $H^*(S_H(\Gamma'), \mathbb{C})$ is non trivial.

By duality we have a map

$$E_c\colon H_c^*(S(\Gamma),\mathbb{C}) \to H^*(\mathfrak{g}, K, \pi).$$

Suppose again that $0 \neq [Q^*\omega] \in H^*(\mathfrak{h}, K \cap H, \pi_H)$.

Conjecture. Suppose that $[Q^* E_c^* \omega] \neq 0$. Then for sufficiently small Γ there exists a "restriction" map

$$\mathrm{res}_c\colon H^*(S(\Gamma),\mathbb{C}) \to H_c^*(S_H(\Gamma'),\mathbb{C})$$

so that $\mathrm{res}_c[\omega] \neq 0$.

If the restriction of π to $\mathfrak{h}$ is a direct sum of unitary irreducible representations, the first conjecture is true.

References

[1] Ash, A., Borel, A., Generalized modular symbols. In *Cohomology of arithmetic groups and automorphic forms*, Lecture Notes in Math. 1447, Springer-Verlag, Berlin 1990, 57–75.

[2] Borel, A., Stable real cohomology of arithmetic groups II. In *Manifolds and Lie groups*, Progr. Math. 14, Birkhäuser, Boston, MA, 1981, 21–55.

[3] Borel, A., Wallach, N., *Continuous Cohomologuy, Discrete Subgroups, and Representations of Reductive Groups*. Ann. of Math. Stud. 94, Princeton University Press, Princeton, NJ, 1980.

[4] Clozel, L., Venkataramana, T. N., Restriction of the cohomology of a Shimura variety to a smaller Shimura variety. *Duke Math. J.* **95** (1) (1998), 51–106.

[5] Franke, J., Harmonic analysis in weighted L^2-spaces. *Ann. Sci. École Norm. Sup.* (4) **31** (2) (1998), 181–279.

[6] Franke, J., A Topological Model for Some Summand of the Eisenstein Cohomology of Congruence Subgroups. Preprint, Bielefeld University, 1991.

[7] Harder, G., General aspects in the theory of modular symbols. In *Séminaire de théorie des nombres* (Paris 1981–82), Progr. Math. 38, Birkhäuser, Boston, MA, 1983, 72–88.

[8] Kudla, S., Derivatives of Eisenstein series and arithmetic geometry. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 173–183.

[9] Kudla, S., Millson, J., Intersection numbers of cycles on locally symmetric spaces and Fourier coefficients of holomorphic modular forms in several complex variables. *Inst. Hautes Études Sci. Publ. Math.* **71** (1990), 121–172.

[10] Langlands, R., *On the functional equations satisfied by Eisenstein series.* Lecture Notes in Math. 544, Springer-Verlag, Berlin 1976.

[11] Mahnkopf, J., Modular symbols and values of L-functions on Gl_3. *J. Reine Angew. Math.* **497** (1998), 91–112.

[12] Mahnkopf, J., Eisenstein cohomology and the construction of p-adic L-functions. *Compositio Math.* **124** (3) (2000), 253–304.

[13] Rohlfs, J., Speh, B., Pseudo-Eisenstein forms and cohomology of arithmetic groups I. *Manuscripta Math.* **106** (4) (2001), 505–518.

[14] Rohlfs, J., Speh, B., Pseudo-Eisenstein forms and cohomology of arithmetic groups II. In *Algebraic groups and arithmetic*, Tata Inst. Fund. Res., Mumbai 2004, 63–89.

[15] Schwermer, J., *Kohomologie arithmetisch definierter Gruppen und Eisensteinreihen.* Lecture Notes in Math. 988, Springer-Verlag, Berlin (1983).

[16] Speh, B., Unitary representations of $\mathrm{GL}(n, \mathbb{R})$ with nontrivial $(\mathfrak{g}, K)$-cohomology. *Invent. Math.* **71** (1983), 443–465.

[17] Speh, B., Venkataramana, T. N., Construction of some generalised modular symbols. *Pure Appl. Math. Q.* **1** (4) (2005), 737–754.

[18] Venkataramana, T. N., Cohomology of compact locally symmetric spaces. *Compositio Math.* **125** (2) (2001), 221–253.

[19] Tong, Y., Wang, S., Geometric realisation of discrete series representations for semisimple symmetric spaces. *Invent. Math.* **96** (1989), 425–458.

Department of Mathematics, Cornell University, Ithaca, NY 14850, U.S.A.
E-mail: speh@math.cornell.edu

Some results on compactifications of semisimple groups

Tonny A. Springer

Abstract. This paper deals with recent results involving a compactification X of a semisimple group G. The emphasis is on the case that G is adjoint and X is its wonderful compactification. Group theoretical constructions in G have repercussions in X. The paper describes a number of them.

Mathematics Subject Classification (2000). Primary 14M17; Secondary 14M15.

Keywords. Semisimple groups, compactifications.

1. Introduction

1.1. Notations. G is a connected semisimple linear algebraic group over the algebraically closed field k of characteristic $p \geq 0$. Fix a maximal torus T of G and a Borel subgroup $B = T.U$, where U is the unipotent subgroup of B. The Weyl group $N_G(T)/T$ is denoted by W. It acts on T.

For $w \in W$ we denote by $\dot{w}$ a representative of $w \in N_G(T)$, not necessarily always the same.

R is the root system of (G, T) and $R^+ \subset R$ the system of positive roots defined by B. Its set of simple roots is denoted by I. We identify it with the set of simple reflections. l is the corresponding length function on W.

For $J \subset I$, W_J is the subgroup of W generated by J and W^J the set of minimal length coset representatives for W/W_J. The longest element of W_J is $w_{0,J}$.

1.2. Compactification. $G \times G$ acts on G by $(x, y).z = xyz^{-1}$ $(x, y, z \in G)$. By a *compactification* of G we mean an irreducible normal projective $G \times G$-variety, containing G as an open $G \times G$-stable subvariety. The theory of embeddings of spherical varieties (due to Luna and Vust, see [Kn]) can be applied to the $G \times G$-variety G to analyze such compactifications.

We shall be concerned here mainly with the particular case that G is adjoint and X is the wonderful compactification of G. This was first constructed for $k = \mathbb{C}$ by De Concini and Procesi in [DP1]. The case of an arbitrary algebraically closed k was dealt with in [St]. (See also [DS], where fields of definition are also taken into account).

In the construction given in these papers one uses a suitable finite-dimensional projective representation $\rho\colon G \to \mathrm{PGL}(V)$ of G and X is defined to be the closure in

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

$\mathbb{P}(\mathrm{End}(V))$ of the image $\rho(G)$ (this turns out to be independent of the choice of ρ).

Then X is a compactification of G. It is a smooth projective $G \times G$-variety. A key property is that X contains a unique closed $G \times G$-orbit, isomorphic to $G/B^- \times G/B$, where B^- is the opposite of B (see Lemma 1 below). The complement $X - G$ is a union of smooth divisors D_i indexed by the simple roots $i \in I$, with normal crossings.

For $J \subset I$ let $X_J \subset \bigcap_{i \in I-J} D_i$ be the set of points not lying in a smaller intersection of the same kind. The X_J $(J \subset I)$ are the $G \times G$-orbits in X. Then $X_I = G$, $X_{I-\{i\}}$ is the orbit which is open in D_i and $X_\emptyset$ is the closed orbit.

[DP2] analyzes general smooth compactifications of G. The wonderful one is shown to be "minimal".

There are other constructions of X:

(a) as the closure of the $G \times G$-orbit of the diagonal in the variety of Lie subalgebras of $\mathrm{Lie}(G \times G)$ (see [DP1, sect. 6]),

(b) as the closure of the $G \times G$-orbit of the diagonal of $G/B \times G/B$, viewed as a point of the Hilbert scheme of $G/B \times G/B$ (see [B2]).

If G is arbitrary there does not seem to be a canonical smooth compactification. One can construct a not necessarily smooth one in the following manner.

Let G_{ad} be the corresponding adjoint group and X_{ad} its wonderful compactification. Let X be the normalization of X_{ad} in the function field $k(G)$ (a finite extension of $k(G_{\mathrm{ad}})$).

Then X is a compactification of G. The homomorphism $G \to G_{\mathrm{ad}}$ extends to a $G \times G$-morphism $X \to X_{\mathrm{ad}}$. It induces a bijection of the sets of $G \times G$-orbits.

We recall some facts, to be elaborated on later in the context of compactifications.

1.3. Bruhat's Lemma. G is the disjoint union of the locally closed subsets $G_w = B\dot{w}B$. In other words: $B \times B$ acts on G with finitely many orbits, indexed by the elements of W.

There is an order $\leq$ on W (the Bruhat–Chevalley order) such that for $x, w \in W$ the orbit G_x lies in the closure $\overline{G_w}$ if and only if $x \leq w$.

The flag variety G/B is a smooth projective G-variety. The closed subvarieties $S_w = \overline{G_w}/B$ $(w \in W)$ are the Schubert varieties.

1.4. Conjugation action. Let $G_{\mathrm{d}} \simeq G$ be the diagonal subgroup of $G \times G$. The restriction to G_{d} of the $G \times G$-action on G is the conjugation action of G on itself.

We have a partition of G into G_{d}-stable closed subsets, each of which consists of the elements whose semisimple part lies in a given conjugacy class. We call these subsets *Steinberg fibers*. Regular semisimple conjugacy classes are examples.

The Steinberg fibers are the fibers of a flat morphism $G \to W\backslash T$.

1.5. Character sheaves. There is a geometric character theory for G, embodied in Lusztig's theory of character sheaves (see [L1]). Character sheaves are certain

conjugation-equivariant irreducible perverse sheaves on G. Ingredients in their construction are $B \times B$-equivariant perverse sheaves on G, supported by the closure of some G_w.

2. $B \times B$-action on a compactification

2.1. The $B \times B$-orbits. G is assumed to be adjoint and X is its wonderful compactification. The question of extending Bruhat's Lemma to X, i.e. of describing the $B \times B$-action on X arises naturally. It was studied in [B1], [Sp1] and more recently in [HT2]. The last paper also deals with non-adjoint groups.

We first describe in more detail the $G \times G$-orbits X_J in X.

For $J \subset I$ denote by $P_J \supset B$ the corresponding standard parabolic subgroup and by $P_J^- \supset B^-$ its opposite. We denote by L_J the Levi subgroup of P_J and P_J^- containing T and by G_J the adjoint group of L_J. The image of T in G_J is a maximal torus denoted by T_J.

Let λ be a cocharacter of T (a homomorphism $\lambda\colon \mathbb{G}_m \to T$). Then $\lambda(0) = \lim_{\xi \to 0} \lambda(\xi)$ is a well-defined point of X, as X is complete. λ also defines a linear function on the character group of T, denoted by the same symbol.

The $G \times G$-orbits X_J ($J \subset I$) in X can be described as follows.

Lemma 1. (i) *There is a unique base point $h_J \in X_J$ such that for all cocharacters λ with $\lambda(\alpha) = 0$ for $\alpha \in J$ and $\lambda(\alpha) > 0$ for $\alpha \in I - J$ we have $\lambda(0) = h_J$.*

(ii) *The orbit map $(x, y) \mapsto (x, y).h_J$ induces an isomorphism of the quotient variety $(G \times G) \times_{P_J^- \times P_J} G_J$ onto X_J.*

For (i) see [DS, sect. 3]. The quotient of (ii) is relative to the right action of $P_J^- \times P_J$ on $G \times G$ and the left action on G_J given by $(x, y).z = \bar{x} z \bar{y}^{-1}$, the bars denoting projection on G_J in P_J, respectively P_J^-. See [Sp1, p. 73].

By (ii) the closed orbit $X_\emptyset$ is isomorphic to $G/B^- \times G/B$ (as was already mentioned).

Proposition 1. (i) *A $B \times B$-orbit in X_J is of the form*

$$[J, x, w] = (B \times B).(\dot{x}, \dot{w}).h_J,$$

with unique $w \in W$, $x \in W^J$.

(ii) $\dim[J, x, w] = l(w_{0,J}) - l(x) + l(w) + |J|$.

See [Sp1, p. 74]. The result is due to Brion [B1]. $[I, 1, w]$ is the set G_w of 1.3.

The next result describes the closure relations between the $[I, x, w]$.

Proposition 2. *Let $J, J' \subset I$, $x \in W^J$, $x' \in W^{J'}$, $w, w' \in W$. Then $[J', x', w'] \subset \overline{[J, x, w]}$ if and only if $J' \subset J$ and there exists $u \in W_J$ such that $xu \leq x'$, $w' \leq wu$.*

See [HT2, Prop. 6.3]. [Sp1, Prop. 2.4] gives a somewhat more complicated description.

The closures $\overline{G_w}$ (in X) are the *large Schubert varieties*, first studied in [BP]. The large Schubert variety of minimal dimension is the closure $\overline{B}$. It follows from Proposition 2 that it is the union of the $[J, x, w]$ with $w \leq x$.

In [BP] it is shown how the geometry of $\overline{B}$ can be used to understand a result about the simply connected cover G_{sc} of G, namely van der Kallen's filtration (see [Kal]) of the coordinate algebra $k[B_{\mathrm{sc}}]$ of the preimage B_{sc} of B in G_{sc}.

More generally, for each J there is a unique $B \times B$-orbit of minimal dimension in X_J, viz. $[J, w_{0,I} w_{0,J}, 1]$.

In the present case there is also a version of the Bott–Samelson–Demazure–Hansen variety. To formulate it succinctly write $B \times B = \boldsymbol{B}$ and denote by $\boldsymbol{P}_h$ minimal parabolic subgroups of $G \times G$ containing $\boldsymbol{B}$ (so $\boldsymbol{P}_h/\boldsymbol{B} = \mathbb{P}^1$). For $\boldsymbol{h} = (h_1, \ldots, h_s)$ put

$$Z_{\boldsymbol{h}} = \boldsymbol{B}^s \backslash (\boldsymbol{P}_{h_1} \times \cdots \times \boldsymbol{P}_{h_s} \times \overline{[J, w_{0,I} w_{0,J}, 1]}),$$

the quotient for the $\boldsymbol{B}^s$-action (with obvious notations)

$$(b_1, \ldots, b_s).(p_1, \ldots, p_s, x) = (p_1 b_1^{-1}, b_1 p_2 b_2^{-1}, \ldots, b_{s-1} p_s b_s^{-1}, b_s.x).$$

The $G \times G$-action on X defines a morphism $\phi_{\boldsymbol{h}} \colon Z_{\boldsymbol{h}} \to X$.

Proposition 3. *Given $x \in W^J$, $w \in W$ there exist $\boldsymbol{h}$ such that $\phi_{\boldsymbol{h}}$ is a proper birational morphism of $Z_{\boldsymbol{h}}$ onto $\overline{[J, x, w]}$.*

The proof is along familiar lines, it uses reduced decompositions of w and of $w_{0,J} w_{0,I} x$. However, one cannot claim that $\phi_{\boldsymbol{h}}$ is a resolution, as the varieties $\overline{[J, w_{0,I} w_{0,J}, 1]}$ are usually not smooth.

For example, if G is simple and of rank > 1, $\overline{B}$ is not smooth, see [Sp1, Cor. 4.8]. The following problem arises naturally.

Problem 1. Construct a $B \times B$-equivariant resolution of $\overline{[J, w_{0,I} w_{0,J}, 1]}$, in particular of $\overline{B}$.

Proposition 4. *A large Schubert variety admits a cellular decomposition* (*paving by affine spaces*).

See [Sp1, p. 81]. The cells can be described explicitly, which leads to a description of the cohomology groups of large Schubert varieties. In particular, their odd cohomology vanishes (see [loc. cit., 2.11]).

Problem 2. Do all $B \times B$-orbit closures have cellular decompositions?

Proposition 5. *The odd* (*global*) *intersection cohomology of a $B \times B$-orbit closure vanishes.*

See [loc. cit., Thm. 4.11]. The local intersection cohomology of orbit closures will appear in Section 4.

2.2. Algebro-geometric properties of orbit closures. In this subsection G is an arbitrary semisimple group and X is a compactification of G, as in 1.2. For Frobenius splittings and their various refinements we refer to [BK].

Theorem 1. *Let $p > 0$. X admits a $B \times B$-canonical Frobenius splitting which compatibly splits the closures of all $B \times B$-orbits.*

This is [HT2, Prop. 7.1], where it is deduced from the slightly weaker result in [BK, Thm. 6.1.12].
The theorem has the following corollaries, in any characteristic. They are proved by familiar arguments. Let $Z \subset X$ be a $B \times B$-orbit closure.

Corollary 1. *Let $\mathcal{L}$ be an ample line bundle on Z.*

(i) $H^i(Z, \mathcal{L}) = 0$ *for* $i > 0$.

(ii) *If $Z' \subset Z$ is another orbit closure then the restriction map*

$$H^i(Z, \mathcal{L}) \to H^i(Z', \mathcal{L})$$

is surjective.

Corollary 2. *Z is normal and Cohen–Macaulay.*

In fact, more is proved in [loc. cit.], namely that all $B \times B$-orbit closures are globally F-regular. This property also entails the two Corollaries (and more). We will not go into this.

Corollary 2 was first proved by Brion in [B3].

Let $\mathcal{L}$ be an ample line bundle on X. Chirivì and Maffei in [CM] constructed a "standard monomial basis" of the space of global sections $H^0(X, \mathcal{L})$. K. Appel recently showed (see [A]) that this basis is compatible with the $B \times B$-orbit closures.

3. The G_d-action

In this section G is adjoint and X is its wonderful compactification. This section discusses results about the G_{d}-action on X. Notations are as in Section 2.

If σ is an automorphism of G we have a σ-twisted $G \times G$-action on X: $(x, y).{}_\sigma z = (x, \sigma y).z$. The induced G_{d}-action on G is σ-twisted conjugacy: $(x, y) \mapsto xy(\sigma x)^{-1}$. Several of the results of this section extend to twisted actions.

3.1. A partition of X. The partition to be described is essentially due to Lusztig (see [L2, 12.3]). The present formulation was given by He (in [H1, sect. 2]). A similar result also occurs in [EL].

For $J \subset I$ and $W \in W^J$ put

$$X_{J,w} = G_{\mathrm{d}}.[J, w, 1]. \tag{1}$$

Theorem 2. (i) *X_J is the disjoint union of the $X_{J,w}$ ($w \in W^J$).*

(ii) *$X_{J,w}$ is locally closed and irreducible, of dimension* $\dim G - l(w) - |I - J|$.

(iii) *For $w \in W^J$ there exist a connected reductive group G_w and an automorphism σ_w of it such that there is a bijection of the set of G_{d}-orbits in $X_{J,w}$ onto the set of σ_w-twisted conjugacy classes of G_w.*

The proofs of (i) in [L] and [H1] use a combinatorial machinery. (ii) and (iii) are also due to Lusztig (see [L2, sect. 8]).

Remark. For $J = \emptyset$ part (i) of the theorem is a familiar variant of Bruhat's lemma.

We next describe the closure relations between the $X_{J,w}$, following [H2]. Let $\mathcal{I}$ be the set of pairs (J, x) with $J \subset I,\ x \in W^J$. Define a relation $\leq$ on $\mathcal{I}$ by

$$(J, x) \leq (K, y) \text{ if and only if } J \subset K \text{ and } x \geq z^{-1} y z \text{ for some } z \in W_K.$$

Theorem 3. (i) *$\leq$ defines an order on $\mathcal{I}$.*

(ii) *If $(J, x),\ (K, y) \in \mathcal{I}$ then $X_{J,w} \subset \overline{X_{K,y}}$ if and only if $(J, x) \leq (K, y)$.*

See [H2, sect. 3, 4].

Proposition 6. *If $\overline{X_{K,y}}$ contains only finitely many G_{d}-orbits then it has a cellular decomposition.*

See [loc. cit., sect. 5].

3.2. The closure of Steinberg fibers. Let $F \subset G$ be a Steinberg fiber (see 1.4). Its closure $\overline{F}$ is an irreducible closed G_{d}-stable subset of X. An example is the unipotent variety of X, the closure of the unipotent variety G_u of G.

Lemma 2. *There is $t \in T$ such that $\overline{F} = G_{\mathrm{d}}.t\overline{U}$*

See [Sp2, Lemma 1.4].

This leads to the problem of describing $\overline{U}$. Some partial results are given in [loc. cit.]. They use the fact (a consequence of completeness) that a point of $\overline{U}$ can be obtained by "specializing ξ to 0" from a point of $U(K)$ where $K = k((\xi))$, the field of formal Laurent series.

Let again G_{sc} be the simply connected cover of G, with Borel group $B_{\mathrm{sc}} = T_{\mathrm{sc}}.U_{\mathrm{sc}}$, B_{sc} and T_{sc} lying over B and T. Then $U_{\mathrm{sc}} \simeq U$.

Put

$$H = \{g \in G_{\mathrm{sc}}(k[[\xi]]) \mid g(0) \in B_{\mathrm{sc}}\}.$$

This is the Iwahori subgroup of $G(K)$ defined by B_{sc}. Let W_a be the affine Weyl group associated to T_{sc}. Then we have the Bruhat decomposition $G(K) = HW_aH$. whence a map $G(K) \to W_a$.

Problem 3. Determine the image in W_a of $U_{\mathrm{sc}}(K)$.

A solution of this problem will be useful for describing of $\overline{U}$, see [loc. cit.].

The main fact about the closures $\overline{F}$ is that they all intersect the boundary $X - G$ of X in the same set. More precisely, we have the following result. For $w \in W$ we denote by $\mathrm{supp}(w) \subset I$ the set of simple reflections occurring in a reduced decomposition of w.

Theorem 4.

$$\overline{F} - F = \coprod_{J \neq I} \coprod_{\substack{w \in W^J \\ \mathrm{supp}(w) = I}} X_{J,w}$$

This was first proved by He in [H1, Thms. 4.3, 4.5], via a laborious case by case check. In [HT1] a shorter proof is given and the result is extended to the σ-twisted case.

3.3. We sketch a simplified version of the proof of the theorem. It uses the following steps. F is a Steinberg fiber.

(a) $\overline{F} \cap X_\emptyset \neq \emptyset$.

By Lemma 2 it suffices to show that $\overline{U} \cap X_\emptyset \neq \emptyset$. This follows from the results of [Sp2, sect. 3], for example from [loc. cit., Cor. 3.8] with $w = w_{0,I}$.

(b) If $J \neq I$ and $X_{J,w} \cap \overline{F} \neq \emptyset$ then $\mathrm{supp}(w) = I$.

This is established using an argument from the proof of [H1, Thm. 4.3]. Assume that $i \notin \mathrm{supp}(w)$. Let ϖ_i be the fundamental weight associated to i and let (ρ, V) be an irreducible representation of G with lowest weight $-n\varpi_i$ $(n > 0)$. Then ρ extends to a G_{d}-equivariant morphism $X \to \mathbb{P}(\mathrm{End}(V))$ (see [DS, 3.15]). The image $\rho(\overline{F} - F)$ consists of nilpotent lines in $\mathrm{End}(V)$. On the other hand a lowest weight vector is an eigenvector of $\rho(\dot{w})$ with a nonzero eigenvalue and $\rho(h_J)$ is projection on the line of lowest weight vectors. Using (1) and Proposition 1 (i) one sees that this contradicts nilpotency.

(c) $\overline{F}$ and X_J intersect properly if $J \neq I$.

It suffices to prove this for $J = \emptyset$. From (a) it follows that $\dim \overline{F} \cap X_\emptyset \geq \dim F - |I|$. (b) implies, on the other hand, that the intersection has dimension $\leq \dim F - |I|$.

(d) Let $i \in I$. By (c) $\dim(\overline{F} \cap X_{I-\{i\}}) = \dim F - 1$. By Theorem 2 (i) there must be a set $X_{J,w}$ whose intersection with $\overline{F} \cap X_{I-\{i\}}$ is dense in $X_{I-\{i\}}$. Then $J \subset I - \{i\}$ and

$$\dim G - l(w) - |I - J| = \dim X_{J,w} \geq \dim F - 1 = \dim G - |I| - 1,$$

and $l(w) \leq |J| + 1$. But by (b) $l(w) \geq |I|$. We conclude that $|J| = |I| - 1$ and $l(w) = |I|$. We then must have $J = I - \{i\}$. Moreover, w is a Coxeter element i.e., $\mathrm{supp}(w) = I$ and $l(w) = |I|$.

(e) $W^{I-\{i\}}$ contains a unique Coxeter element c_i.

This is proved by induction on $|I|$.

(f) We conclude from (d) that $\overline{F} \cap X_{I-\{i\}} = \overline{X_{I-\{i\},c_i}}$ for all $i \in I$. Since for $J \neq I$

$$X_J = \bigcap_{i \notin J} X_{I-\{i\}}$$

this implies that the intersection $\overline{F} \cap (X - G)$ is independent of F. With a little more work the theorem follows.

3.4. Algebro-geometric properties. Let X_{sc} be a compactification of the simply connected cover G_{sc}. Let X_i $(1 \leq i \leq n)$ be the irreducible components of $X_{\mathrm{sc}} - G_{\mathrm{sc}}$. They all have codimension 1.

Proposition 7. *Let $p > 0$. Let F be a Steinberg fiber in G_{sc}.*

(i) *X_{sc} admits a Frobenius splitting which compatibly splits $\overline{F}$ and the X_i $(1 \leq i \leq n)$.*

(ii) *$\overline{F}$ is normal and Cohen–Macaulay.*

For (a somewhat stronger version of) (i) see [T, Thm. 8.2] and for (ii) [loc. cit., Thm. 10.2].

Notice that this result covers the wonderful compactification X of G if G has trivial center. For arbitrary adjoint G a partial result is proved in [LT]. For $i \in I$ let χ_i be the fundamental character of G_{sc} associated to i. Put

$$\tilde{F}_0 = \{g \in G_{\mathrm{sc}} \mid \chi_i(g) = 0 \text{ for all } i\},$$

this is a Steinberg fiber in G_{sc}. Its image in G is a Steinberg fiber F_0 in G, the zero fiber.

Proposition 8. *Let $p > 0$. X admits a Frobenius splitting which compatibly splits F_0 and the components of $X - G$.*

See [loc. cit., Thm. 8.1]. It is also pointed out that the result cannot be true for arbitrary Steinberg fibers in G.

Remark. The appearance of the zero fiber is somewhat curious. Over $\mathbb{C}$ it appears in [Ka] in another context. I learned from J.-P. Serre (private communication) that for a quasi-simple group over $\mathbb{C}$, F_0 has been determined (case by case). Its elements are regular and have finite order. The characteristic p case does not seem to have been analyzed.

Problem 4. $(p > 0)$ Does X admit a Frobenius splitting which compatibly splits an arbitrary $\overline{F}$?

Problem 5. Is $\overline{F}$ normal and Cohen–Macaulay?

Problem 6. Does $\overline{F}$ admit a cellular decomposition?

An example given in [Sp2, 4.3] with $G = \mathrm{PGL}_3$ shows that $\overline{F}$ need not be smooth.

Problem 7. Determine the intersection cohomology (local and global) of $\overline{F}$.

4. Character sheaves on X

In this section G is an adjoint group and X is its wonderful compactification.

4.1. $B \times B$-equivariant perverse sheaves. The definition of character sheaves on X uses certain $B \times B$-equivariant perverse sheaves on X, which we first have to introduce.

If S is a torus let $C(S)$ be its character group and put

$$\hat{C}(S) = C(S) \otimes (\mathbb{Z}_{(p)}/\mathbb{Z}),$$

where $\mathbb{Z}_{(p)}$ is the localization of $\mathbb{Z}$ at the prime ideal (p). The elements of $\hat{C}(S)$ parametrize tame rank one local systems on S (also called Kummer local systems). We work in l-adic cohomology, with a coefficient field E (e.g. $\overline{\mathbb{Q}_l}$).

Let $v = [J, x, w]$ be a $B \times B$-orbit in X, as in Proposition 1. Using that it is a homogeneous space for $B \times B$ one constructs a morphism $\phi\colon v \to T_J$, where T_J is the maximal torus of G_J of 2.1 (see [H3, 3.1]). For $\xi \in \hat{C}(T_J)$ we have a local system $\mathcal{L}_{\xi,v} = \phi^*\xi$ on v. Let $\mathcal{I}_{\xi,v}$ be its perverse extension, a perverse sheaf on X (for l-adic cohomology) supported by $\bar{v}$, whose restriction to v is $\mathcal{L}_{\xi,v}[\dim v]$.

[Sp1, sect. 5] deals with these perverse sheaves. It is shown that they are even, i.e. that their cohomology sheaves are zero in dimensions $\not\equiv \dim v \pmod 2$.

In the next lemma one uses that $\hat{C}(T_J)$ can be viewed as a subset of $\hat{C}(T)$ (see [loc. cit., 1.7]) and that W acts on $\hat{C}(T)$.

Lemma 3. *If $x.\xi = w.\xi$ then $\mathcal{I}_{\xi,v}$ is a B_{d}-equivariant irreducible perverse sheaf on X.*

See [H3, 3.1].

4.2. Character sheaves. Character sheaves on a reductive group were introduced by G. Lusztig in the 1980s, in a long series of papers. [L1] gives a brief exposition of the results of these papers. [MS] is a report on part of the results. The definition of character sheaves used there is slightly different from Lusztig's.

In [L2] Lusztig defines character sheaves on the compactification X. I noticed (unpublished) that the approach of [MS] could also be followed to do this. But it is not obvious that the two definitions of character sheaves on X are equivalent.

Independently, Xuhua He also came to the definition based on [MS]. He proved in [H3] the equivalence with Lusztig's definition. I shall not go into Lusztig's definition. I will only report on the other one.

B acts on $G \times X$ by $b.(g, x) = (gb^{-1}, (b, b).x)$. Let $G \times_B X$ be the quotient and $\alpha\colon G \times X \to G \times_B X$ the quotient map. The G_{d}-action on X induces a proper morphism $\mu\colon G \times_B X \to X$.

Let $\mathcal{I}_{\xi,v}$ be as in Lemma 3. Then $A = E[\dim G] \boxtimes \mathcal{I}_{\xi,v}$ is an irreducible perverse sheaf on $G \times X$ and there is an irreducible perverse sheaf $\tilde{A}$ on $G \times_B X$ with $A = \alpha^*\tilde{A}$.

Put $C_{\xi,v} = \mu_! \tilde{A}$. By the decomposition theorem this is a semisimple complex on X, i.e. a direct sum of shifted irreducible perverse sheaves on X. The perverse sheaves occurring in the $C_{\xi,v}$ (if ξ and v vary) are the character sheaves on X. They are G_{d}-equivariant.

The nonzero restrictions of these character sheaves to the open subvariety G of X are Lusztig's original character sheaves. More generally, for $J \subset D$ we call character sheaf on X_J the restriction to X_J of a character sheaf on X which is obtained as above from an orbit $v \subset X_J$.

The character sheaves on X deserve a further study. Here are a few problems.

Problem 8. Analyze the restriction of a character sheaf on X to a $G \times G$-orbit X_J. Can such restrictions be described in terms of character sheaves on X_J?

Problem 9. Are character sheaves on X even?

4.3. Finite ground fields. Now let k be an algebraic closure of the finite field $\mathbb{F}_q$ and let $F\colon a \mapsto a^q$ be the Frobenius automorphism of k. Assume that G is defined over k. Then so is X, by [DS, Prop. 3.11].

Let A be a character sheaf on G whose support is not contained in $X - G$. The restriction of A to G is a character sheaf on G. Assume that A "comes from $\mathbb{F}_q$", meaning that $F^*A \simeq A$. Fix an isomorphism $\phi\colon F^*A \simeq A$. It can be normalized such as to be unique up to a root unity (see [L1, p. 178]).

$x \in X^F = X(\mathbb{F}_q)$ being an $\mathbb{F}_q$-rational point of X, ϕ defines linear maps ϕ_x^i of the cohomology stalks $H^i(A)_x$. Define a function χ_ϕ on X^F by

$$\chi_\phi(x) = \Sigma_i (-1)^i \operatorname{Tr}(\phi_x^i, H^i(A)_x).$$

For $x \in G$ one obtains a class function on the finite Lie group G^F, which can be viewed as a generalized character of G^F. This function on G^F has boundary values, viz. the values of χ_ϕ on points of $X^F - G^F$.

Problem 10. Can one define and compute boundary values of irreducible characters of G^F?

Acknowledgement. I am grateful to M. Brion for his comments on this paper.

References

[A] Appel, K., Standard monomials for wonderful group compactifications. Preprint math. RT/0512020.

[B1] Brion, M., The behaviour at infinity of the Bruhat decomposition. *Comment. Math. Helv.* **71** (1998), 137–174.

[B2] Brion, M., Group completions via Hilbert schemes. *J. Algebraic Geom.* **13** (2003), 603–626.

[B3] Brion, M., Multiplicity free subvarieties of flag varieties. In *Commutative algebra* (Grenoble/Lyon, 2001), Contemp. Math. 331, Amer. Math. Soc., Providence, RI, 2003, 13–23.

[BK] Brion, M., and Kumar, S., *Frobenius Splitting Methods in Geometry and Representation Theory*. Progr. Math. 231, Birkhäuser, Boston, MA, 2004.

[BP] Brion, M., and Polo, P., Large Schubert varieties. *Represent. Theory* **4** (2000), 97–126.

[CM] Chirivì, R., and Maffei, A., The ring of sections of a complete symmetric variety. *J. Algebra* **261** (2003), 310–326.

[DP1] De Concini, C., and Procesi, C., Complete symmetric varieties. In *Invariant Theory* (Montecatini, 1982), Lecture Notes in Math. 996, Springer-Verlag, Berlin 1983, 1–44.

[DP2] De Concini, C., and Procesi, C., Complete symmetric varieties II, Intersection theory. In *Algebraic groups and related topics* (Kyoto/Nagoya, 1983), Adv. Stud. Pure Math. 6, Kinokuniya/North-Holland, Amsterdam 1985, 481–513.

[DS] De Concini, C., and Springer, T. A., Compactification of symmetric varieties. *Transform. Groups* **4** (1999), 273–300.

[EL] Evens, S., and Lu, J.-H., On the variety of Lagrangian subalgebras, II. Preprint math. QA/0409236.

[H1] He, X., Unipotent variety in the group compactification. *Adv. in Math.* **203** (1) (2006), 109–131.

[H2] He, X., The G-stable pieces of the wonderful compactification. *Trans. Amer. Math. Soc.*, to appear.

[H3] He, X., The character sheaves on the group compactification. *Adv. in Math*, to appear.

[HT1] He, X., and Thomsen, J. F., Closures of Steinberg fibers in twisted wonderful compactifications. Preprint math. AG/0506087.

[HT2] He, X., and Thomsen, J. F., Geometry of $B \times B$-orbit closures in equivariant embeddings. Preprint math. AG/0510088.

[Ka] Kac, V., Simple Lie groups and the Legendre symbol. In *Algebra Carbondale 1980*, Lecture Notes in Math. 848, Springer-Verlag, Berlin 1981, 110–123.

[Kal] van der Kallen, W., Longest weight vectors and excellent filtrations. *Math. Z.* **201** (1989), 19–31.

[Kn] Knop, F., The Luna-Vust theory of spherical embeddings. In *Proc. Hyderabad Conf. on Algebraic Groups*, Manoj Prakashan, Madras 1991, 225–248.

[L1] Lusztig, G., Introduction to character sheaves. In *The Arcata Conference on Representations of Finite Groups* (Arcata, Calif., 1986), Proc. Symp. Pure Math. 47 (part 1), Amer. Math. Soc., Providence, RI, 1987, 165–179.

[L2] Lusztig, G., Parabolic character sheaves I. *Moscow Math. J.* **4** (2004), 153–179; II, ibid., 869–896.

[LT] Lynderup, T. H., and Thomsen, J. F., On compactifications of the Steinberg zero-fiber. Preprint math. AG/0506348.

[MS] Mars, J. G. M., and Springer, T. A., Character sheaves. In *Orbites unipotentes et représentations III*, *Astérisque* **173–174** (1989), 111–198.

[Sp1] Springer, T. A., Intersection cohomology of $B \times B$-orbit closures in group compactifications. *J. Algebra* **258** (2002), 71–111.

[Sp2] Springer, T. A., Some subvarieties of a group compactification. In *Proc. Conf. on Algebraic Groups* (Mumbai 2004), to appear.

[St] Strickland, E., A vanishing theorem for group compactifications. *Math. Ann.* **277** (1987), 165–171.

[T] Thomsen, J. F., Frobenius splitting of equivariant closures of regular conjugacy classses. Peprint math.AG/0502114.

Mathematisch Instituut, Budapestlaan 6, 3584 CD Utrecht, The Netherlands
E-mail: springer@math.uu.nl

Quasiconformal geometry of fractals

Mario Bonk*

Abstract. Many questions in analysis and geometry lead to problems of quasiconformal geometry on non-smooth or fractal spaces. For example, there is a close relation of this subject to the problem of characterizing fundamental groups of hyperbolic 3-orbifolds or to Thurston's characterization of rational functions with finite post-critical set.

In recent years, the classical theory of quasiconformal maps between Euclidean spaces has been successfully extended to more general settings and powerful tools have become available. Fractal 2-spheres or Sierpiński carpets are typical spaces for which this deeper understanding of their quasiconformal geometry is particularly relevant and interesting.

Mathematics Subject Classification (2000). Primary 30C65; Secondary 20F67.

Keywords. Quasiconformal maps, analysis on fractals.

1. Introduction

A homeomorphism on $\mathbb{R}^n$ is called *quasiconformal* if it maps infinitesimal balls to infinitesimal ellipsoids with uniformly controlled eccentricity. While in its substance this notion (for $n = 2$) was introduced by Grötzsch in the late 1920s (see [Kü] for an account of Grötzsch's work), the term "quasiconformal" was first used by Ahlfors in 1935 [Ah, p. 213 and p. 242].

The importance of planar quasiconformal mappings was only fully realized after Teichmüller had published his groundbreaking work on the classical moduli problem for Riemann surfaces around 1940. It took another two decades after the foundational issues of the theory of quasiconformal mappings had been clarified. A great subtlety here is what a priori smoothness assumption to impose on a quasiconformal map in order to get a theory with the desired compactness properties under limits. It turned out that some requirement of Sobolev regularity is appropriate.

Nowadays planar quasiconformal maps are recognized as a standard tool in various areas of complex analysis such as Teichmüller theory, Kleinian groups, and complex dynamics. One of the main reasons for this is that in the plane a flexible existence theorem for quasiconformal maps is available in the Measurable Riemann Mapping Theorem.

*Supported by NSF grants DMS-0200566 and DMS-0244421. The author would like to thank J. Heinonen, B. Kleiner and S. Merenkov for useful discussions and help with this survey.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

In higher dimensions there is no such existence theorem putting severe limitations to the theory. Accordingly, quasiconformal maps for $n \geq 3$ initially had a less profound impact than their planar relatives. This situation changed when Mostow used the theory of higher-dimensional quasiconformal maps in the proof of his celebrated rigidity theorems for rank-one symmetric spaces [Mo]. In this context it also became desirable to extend the classical theory of quasiconformal mappings on $\mathbb{R}^n$ to other settings such as Heisenberg groups which arise as boundaries of complex hyperbolic spaces [KR1], [KR2], [Pa1]. A continuation of this trend was a theory of quasiconformal maps in a general metric space context and recently led to Heinonen and Koskela's theory of quasiconformal maps on Loewner spaces and spaces satisfying a Poincaré inequality [HK].

In this survey we will focus on questions of quasiconformal geometry of low-dimensional fractals such as 2-spheres and Sierpiński carpets. In particular, we will discuss uniformization and rigidity theorems that are motivated by questions in geometric group theory and complex dynamics. For some additional topics not discussed here, we refer to B. Kleiner's article in these conference proceedings.

2. Quasiconformal and quasisymmetric maps

We first give a precise definition of various classes of maps that are related to the concept of quasiconformality (see [BI], [He], [Vä1] for more details).

A homeomorphism $f\colon \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 2$, is called *quasiconformal*, if f belongs to the Sobolev space $W^{1,n}_{\mathrm{loc}}(\mathbb{R}^n, \mathbb{R}^n)$ and if there exists a constant $K \geq 1$ such that the inequality

$$\|Df(x)\|^n \leq K|J_f(x)| \tag{1}$$

is valid for almost every $x \in \mathbb{R}^n$. Here $\|Df(x)\|$ is the norm of the formal differential Df and $J_f = \det(Df)$ is the Jacobian determinant of f. If we want to emphasize K, then we say that f is *K-quasiconformal*. We will use similar language below for concepts that depend on parameters. We exclude $n = 1$ in the following, because in this case the theory of quasiconformal maps has somewhat different features than for $n \geq 2$. An important fact about quasiconformal maps is that they are differentiable almost everywhere. Hence we can replace the formal quantity Df in (1) by the classical derivative.

The "analytic" definition of quasiconformality above applies to more general smooth settings, for example, if f is a map between Riemannian n-manifolds. If one drops the requirement that f is a homeomorphism and allows branching, then one is led to the concept of a *quasiregular* map [Re], [Ri].

A general definition of a quasiconformal map in a metric space context (the "metric" definition) can be given as follows. Suppose $f\colon X \to Y$ is a homeomorphism between metric spaces (X, d_X) and (Y, d_Y). Then f is *quasiconformal* if there exists

a constant $H \geq 1$ such that

$$H_f(x) := \limsup_{r \to 0^+} \frac{\sup\{d_Y(f(x'), f(x)) : d_X(x, x') \leq r\}}{\inf\{d_Y(f(x'), f(x)) : d_X(x, x') \geq r\}} \leq H$$

for all $x \in X$. In $\mathbb{R}^n$, $n \geq 2$, this is equivalent to the analytic definition. In general, this concept is rather weak and does not lead to a useful theory, if one does not impose further restrictions on the spaces X and Y.

A related, but stronger concept is the notion of a quasisymmetric map [TV]. By definition a homeomorphism $f\colon X \to Y$ is called *quasisymmetric*, if there exists a homeomorphism $\eta\colon [0, \infty) \to [0, \infty)$ (which plays the role of a distortion function) such that

$$\frac{d_Y(f(x), f(y))}{d_Y(f(x), f(z))} \leq \eta\left(\frac{d_X(x, y)}{d_X(x, z)}\right),$$

whenever $x, y, z \in X$, $x \neq z$.

The geometric meaning of this condition is that balls are mapped to "round" sets with quantitative control for their "eccentricity". This is a global version of the geometric property of a quasiconformal map. In $\mathbb{R}^n$, $n \geq 2$, a map is quasiconformal if and only if it is quasisymmetric. Inverse maps or compositions of quasisymmetric maps are again quasisymmetric. So the quasisymmetric maps $f\colon X \to X$ on a metric space X form a group that we denote by $\mathrm{QS}(X)$.

In particular when a group action is present, the concept of a quasi-Möbius map is often more suitable than that of a quasisymmetry. Here a homeomorphism $f\colon X \to Y$ is called a *quasi-Möbius map* [Vä2] if there exists a homeomorphism $\eta\colon [0, \infty) \to [0, \infty)$ such that for every 4-tuple (x_1, x_2, x_3, x_4) of distinct points in X, we have the inequality

$$[f(x_1), f(x_2), f(x_3), f(x_4)] \leq \eta([x_1, x_2, x_3, x_4])$$

for the *metric cross-ratio*

$$[z_1, z_2, z_3, z_4] = \frac{d_X(z_1, z_3) d_X(z_2, z_4)}{d_X(z_1, z_4) d_X(z_2, z_3)}.$$

Every quasisymmetric map $f\colon X \to Y$ between metric spaces is also quasi-Möbius. This statement is "quantitative" in the following sense: If f is η-quasisymmetric, then f is $\tilde{\eta}$-quasi-Möbius with $\tilde{\eta}$ only depending on η. Conversely, every quasi-Möbius map between bounded metric spaces is quasisymmetric, but in contrast to the other implication this statement is not quantitative in general. One can also get a quantitative statement here if one assumes that the spaces are bounded and adds a normalization condition for the map.

There is a third way to characterize quasiconformal maps on $\mathbb{R}^n$ based on the concept of the modulus of a path family. We first review some definitions related to this.

Suppose (X, d, μ) is a metric measure space, i.e., (X, d) is a metric space and μ a Borel measure on X. Moreover, we assume that (X, d) is locally compact and that μ is locally finite and has dense support.

The space (X, d, μ) is called (*Ahlfors*) Q*-regular*, $Q > 0$, if the measure μ satisfies

$$C^{-1} R^Q \le \mu(\bar{B}(a, R)) \le C R^Q$$

for each closed ball $\bar{B}(a, R)$ of radius $0 < R \le \operatorname{diam}(X)$ and for some constant $C \ge 1$ independent of the ball. If the measure is not specified, then it is understood that μ is Q-dimensional Hausdorff measure. Note that a complete Ahlfors regular space X is *proper*, i.e., closed balls in X are compact.

A *density* on X is a Borel function $\rho\colon X \to [0, \infty]$. A density ρ is called *admissible* for a path family Γ in X, if

$$\int_\gamma \rho \, ds \ge 1$$

for each locally rectifiable path $\gamma \in \Gamma$. Here integration is with respect to arclength on γ.

If $Q \ge 1$, the Q*-modulus* of a family Γ of paths in X is the number

$$\operatorname{Mod}_Q(\Gamma) = \inf \int \rho^Q \, d\mu,$$

where the infimum is taken over all densities $\rho\colon X \to [0, \infty]$ that are admissible for Γ. If E and F are subsets of X with positive diameter, we denote by

$$\Delta(E, F) = \frac{\operatorname{dist}(E, F)}{\min\{\operatorname{diam}(E), \operatorname{diam}(F)\}}$$

the *relative distance* of E and F, and by $\Gamma(E, F)$ the family of all paths in X connecting E and F.

Suppose (X, d, μ) is a connected metric measure space. Then X is called a Q-*Loewner space*, $Q \ge 1$, if there exists a positive decreasing function $\Psi\colon (0, \infty) \to (0, \infty)$ such that

$$\operatorname{Mod}_Q(\Gamma(E, F)) \ge \Psi(\Delta(E, F)), \tag{2}$$

whenever E and F are disjoint continua in X.

Condition (2) axiomatizes a property of n-modulus in $\mathbb{R}^n$ relevant for quasiconformal geometry to a general metric space setting. Examples for Loewner spaces are $\mathbb{R}^n$, all compact Riemannian manifolds, Carnot groups (such as the Heisenberg group), and boundaries of some Fuchsian buildings. Equivalent to the Loewner property of a space is that it satisfies a Poincaré inequality (see [HK], [He] for more discussion on these topics).

The following theorem is a combination of results by Heinonen and Koskela [HK] and by Tyson [Ty], and characterizes quasiconformal maps on Loewner spaces by a distortion property for modulus.

Theorem 2.1. *Let X and Y be Q-regular Q-Loewner spaces, $Q > 1$, and $f : X \to Y$ a homeomorphism. Then f is quasiconformal if and only if there exists a constant $K \geq 1$ such that*

$$\frac{1}{K} \operatorname{Mod}_Q(\Gamma) \leq \operatorname{Mod}_Q(f(\Gamma)) \leq K \operatorname{Mod}_Q(\Gamma) \tag{3}$$

for every family Γ of paths in X, where $f(\Gamma) = \{f \circ \gamma : \gamma \in \Gamma\}$.

In $\mathbb{R}^n$ this characterization of a quasiconformal map is known as the "geometric" definition. The reason for this terminology is that (3) immediately leads to strong geometric consequences. For example, using this one can show that f is quasisymmetric.

The fact that the analytic, metric, and geometric definitions of a quasiconformal map on $\mathbb{R}^n$ are (quantitatively) equivalent is a rather deep fact. This has recently been generalized to the general Loewner space setting [HKST].

3. The quasisymmetric uniformization problem

The classical uniformization theorem implies that every Riemann surface is conformally equivalent to a "standard" surface carrying a metric of constant curvature. One can ask whether a general metric space version of this fact is true where the class of conformal maps is replaced by the more flexible class of quasisymmetric homeomorphisms.

Quasisymmetric uniformization problem. Suppose X is a metric space homeomorphic to some "standard" metric space Y. When is X quasisymmetrically equivalent to Y?

Here we call two metric spaces X and Y *quasisymmetrically equivalent* if there exists a quasisymmetric homeomorphism from X onto Y. Equivalently, one may ask for a quasisymmetric characterization of Y. Of course, it depends on the context how the term "standard" is precisely interpreted. This general question is motivated by problems in geometric group theory, for example (see Sections 5 and 7). One can also pose a similar uniformization problem for other classes of maps, for example bi-Lipschitz maps (recall that a homeomorphism between metric spaces is called *bi-Lipschitz* if it distort distances by at most a fixed multiplicative amount).

The prime instance for a quasisymmetric uniformization result is the following theorem due to Tukia and Väisälä that characterizes *quasicircles*, i.e., quasisymmetric images of the unit circle $\mathbb{S}^1$ [TV].

Theorem 3.1. *Let X be a metric space homeomorphic to $\mathbb{S}^1$. Then X is quasisymmetrically equivalent to $\mathbb{S}^1$ if and only if X is doubling and linearly locally connected.*

Here a metric space X is called *doubling* if there exists a number N such that every ball in X of radius R can be covered by N balls of radius $R/2$. A metric space X

is called *linearly locally connected* if there exists a constant $\lambda \geq 1$ satisfying the following conditions: If $B(a, r)$ is an open ball in X and $x, y \in B(a, r)$, then there exists a continuum in $B(a, \lambda r)$ connecting x and y. Moreover, if $x, y \in X \setminus B(a, r)$, then there exists a continuum in $X \setminus B(a, r/\lambda)$ connecting x and y. The geometric significance of this condition is that it rules out "cusps" of the space.

A quasisymmetric characterization of the standard $1/3$-Cantor set can be found in [DS]. Work by Semmes [Se1], [Se2] shows that the quasisymmetric characterization of $\mathbb{R}^n$ or the standard sphere $\mathbb{S}^n$ for $n \geq 3$ is a problem that seems to be beyond reach at the moment. The intermediate case $n = 2$ is particularly interesting.

For a metric 2-sphere X to be quasisymmetrically equivalent to the standard 2-sphere $\mathbb{S}^2$ it is necessary that X is linearly locally connected. This alone is not sufficient, but will be if a mass bound assumption is added [BK1].

Theorem 3.2. *Suppose X is a metric space homeomorphic to $\mathbb{S}^2$. If X is linearly locally connected and Ahlfors 2-regular, then X is quasisymmetrically equivalent to $\mathbb{S}^2$.*

This answers a question by Heinonen and Semmes [HS]. A similar result for other simply connected surfaces has been obtained by K. Wildrick [Wi].

The proof of Theorem 3.2 uses approximations of X by graphs that are combinatorially equivalent to triangulations of $\mathbb{S}^2$. Realizing such a triangulation as an incidence graph of a circle packing on $\mathbb{S}^2$, one finds a mapping of X to $\mathbb{S}^2$ on a coarse scale. The main difficulty is to show the subconvergence of this procedure. For this one controls the quasisymmetric distortion by modulus estimates. Incidentally, a very similar algorithm has recently been used to obtain mappings of the surface of the human brain [H-R].

The assumption of Ahlfors regularity for some exponent $Q \geq 2$ is quite natural, because it is satisfied in many interesting cases (for boundaries of Gromov hyperbolic groups for example; see Section 4). There are metric 2-spheres X though that are linearly locally connected and Q-regular with $Q > 2$, but are not quasisymmetrically equivalent to $\mathbb{S}^2$ [Vä3].

The following result can be proved similarly as Theorem 3.2 [BK1].

Theorem 3.3. *Let $Q \geq 2$ and Z be an Ahlfors Q-regular metric space homeomorphic to $\mathbb{S}^2$. If Z is Q-Loewner, then $Q = 2$ and Z is quasisymmetrically equivalent to $\mathbb{S}^2$.*

Note that an analog of this is false in higher dimensions. For example, one can equip $\mathbb{S}^3$ with a Carnot–Carathéodory metric d modeled on the geometry of the Heisenberg group. Then $(\mathbb{S}^3, d)$ is 4-regular and 4-Loewner, but not quasisymmetrically equivalent to standard $\mathbb{S}^3$ [Se1].

In [BK1] a necessary and sufficient condition was established for a metric 2-sphere to be quasisymmetrically equivalent to $\mathbb{S}^2$. This condition is in terms of the behavior of some combinatorially defined moduli of ring domains and too technical to be stated here. The usefulness of any such characterization depends on whether one

can verify its hypotheses in concrete situations such as for fractal 2-spheres coming from dynamical systems as considered in the following Sections 5 and 6.

4. Gromov hyperbolic spaces and quasisymmetric maps

As a preparation for the next sections, we quickly review some standard material on Gromov hyperbolic spaces and groups [GH], [Gr].

Let (X, d) be a metric space, and $\delta \geq 0$. Then X is called *δ-hyperbolic*, if the inequality

$$(x \cdot z)_p \geq \min\{(x \cdot y)_p, (y \cdot z)_p\} - \delta$$

holds for all $x, y, z, p \in X$, where

$$(u \cdot v)_p = \frac{1}{2}\big(d(u, p) + d(v, p) - d(u, v)\big)$$

for $u, v \in X$. If the space X is *geodesic* (this means that any two point in X can be joined by a path whose length is equal to the distance of the points), then the δ-hyperbolicity of X is equivalent to a thinness condition for geodesic triangles. We say that X is *Gromov hyperbolic* if X is δ-hyperbolic for some $\delta \geq 0$. Roughly speaking, this requires that the space is "negatively curved" on large scales. Examples for such spaces are simplicial trees, or Cartan–Hadamard manifolds with a negative upper curvature bound such as the real hyperbolic spaces $\mathbb{H}^n$, $n \geq 2$.

To each Gromov hyperbolic space one can associate a boundary at infinity $\partial_\infty X$ as follows. Fix a basepoint $p \in X$, and consider sequences of points $\{x_i\}$ in X *converging to infinity* in the sense that

$$\lim_{i,j\to\infty} (x_i \cdot x_j)_p = \infty.$$

We declare two such sequences $\{x_i\}$ and $\{y_i\}$ in X as *equivalent* if

$$\lim_{i\to\infty} (x_i \cdot y_i)_p = \infty.$$

Now the *boundary at infinity* $\partial_\infty X$ is defined as the set of equivalence classes of sequences converging to infinity. It is easy to see that the choice of the basepoint p does not matter here. If X is in addition proper and geodesic, then there is an equivalent definition of $\partial_\infty X$ as the set of equivalence classes of geodesic rays emanating from the basepoint p. One declares two such rays as equivalent if they stay within bounded Hausdorff distance. Intuitively, a ray represents its "endpoint" on $\partial_\infty X$. If $\mathbb{H}^n$ is given by the unit ball model, then from this point of view it is clear that $\partial_\infty \mathbb{H}^n = \mathbb{S}^{n-1}$.

The boundary $\partial_\infty X$ comes naturally equipped with a class of "visual" metrics. By definition a metric ρ on $\partial_\infty X$ is called *visual* if there exist $p \in X$, $C \geq 1$, and $\varepsilon > 0$ such that

$$\frac{1}{C} \exp(-\varepsilon(a \cdot b)_p) \leq \rho(a, b) \leq C \exp(-\varepsilon(a \cdot b)_p) \tag{4}$$

for all $a, b \in \partial_\infty X$. In this inequality we used the fact that a "product" $(a \cdot b)_p \in [0, \infty]$ can also be defined for points $a, b \in \partial_\infty X$ in a natural way. Here we have $(a \cdot b)_p = \infty$ if and only if $a = b \in \partial_\infty X$. If X is δ-hyperbolic, then there exists a visual metric ρ with parameter ε if $\varepsilon > 0$ is small enough depending on δ.

In the following we always think of $\partial_\infty X$ as a metric space by equipping it with a fixed visual metric. If ρ_1 and ρ_2 are two visual metrics on $\partial_\infty X$, then the identity map is a quasisymmetric map between (X, ρ_1) and (X, ρ_2) (the visual metrics form a *conformal gauge* – see [He, Ch. 15] for this terminology and further discussion). So the ambiguity of the visual metric is irrelevant if one wants to speak of quasisymmetric maps on $\partial_\infty X$. One should consider the space $\partial_\infty X$ equipped with such a visual metric ρ as very "fractal". For example, if the parameter ε in (4) is very small, then $(\partial_\infty X, \rho)$ will not contain any non-constant rectifiable curves.

Suppose (X, d_X) and (Y, d_Y) are metric spaces. A map $f\colon X \to Y$ is called a *quasi-isometry* if there exist constants $\lambda \geq 1$ and $k \geq 0$ such that

$$\frac{1}{\lambda} d_X(u, v) - k \leq d_Y(f(u), f(v)) \leq \lambda d_X(u, v) + k$$

for all $u, v \in X$ and if

$$\inf_{x \in X} d_Y(f(x), y) \leq k$$

for all $y \in Y$. The spaces X and Y are called *quasi-isometric* if there exists a quasi-isometry $f\colon X \to Y$.

Quasi-isometries form a natural class of maps in the theory of Gromov hyperbolic spaces. For example, Gromov hyperbolicity of geodesic metric spaces is invariant under quasi-isometries. The following fact links these concepts to quasisymmetric maps (see [BS] for more on this subject).

Proposition 4.1. *Let X and Y be proper and geodesic Gromov hyperbolic spaces. Then every quasi-isometry $f\colon X \to Y$ induces a natural quasisymmetric boundary map $\tilde{f}\colon \partial_\infty X \to \partial_\infty Y$.*

The boundary map $\tilde{f}$ is defined by assigning to a point $a \in \partial_\infty X$ represented by the sequence $\{x_i\}$ the point $b \in \partial_\infty Y$ represented by the sequence $\{f(x_i)\}$.

This proposition constitutes a core element in Mostow's proof of rigidity of rank-one symmetric spaces. The point is that a quasi-isometry may locally exhibit very irregular behavior, but gives rise to a quasisymmetric boundary map that can be analyzed by analytic tools.

Suppose G is a finitely generated group, and let S be a set of generators. We will always assume that S is finite and *symmetric* in the sense that if $s \in S$, then $s^{-1} \in S$. The *Cayley graph* $C(G, S)$ of G with respect to S is a graph that has G as its set of vertices. Moreover, we connect two distinct vertices $x, y \in G$ by an edge if there exists $s \in S$ such that $y = xs$. The graph $C(G, S)$ carries a natural path metric that assigns length 1 to each edge. By considering its Cayley graph, one can study properties of the group from a geometric point of view.

The group G is called *Gromov hyperbolic* if $C(G, S)$ is Gromov hyperbolic for some set S of generators of G. If this is the case, then $C(G, S')$ is Gromov hyperbolic for all generating sets S'. This essentially follows from the fact that $C(G, S)$ and $C(G, S')$ are bi-Lipschitz equivalent (and in particular quasi-isometric). This also implies that if we define $\partial_\infty G := \partial_\infty C(G, S)$, then by Proposition 4.1 the boundary at infinity of G is well-defined up to quasisymmetric equivalence. Examples of Gromov hyperbolic groups are free groups and fundamental groups of negatively curved manifolds.

Letting a group element $g \in G$ act on the vertices of $C(G, S)$ by left-translation, we get a natural action $G \curvearrowright C(G, S)$ by isometries. According to Proposition 4.1 this induces a group action $G \curvearrowright \partial_\infty G$, where each group element acts as a quasisymmetric map. In general the distortion function η will be different for different group elements. This changes if one uses the concept of quasi-Möbius maps. In this case the action $G \curvearrowright \partial_\infty G$ is *uniformly quasi-Möbius*, i.e., there exists a distortion function η such that every element $g \in G$ acts as an η-quasi-Möbius map on $\partial_\infty G$ [Pau].

Another important property of the action $G \curvearrowright \partial_\infty G$ is its "cocompactness on triples". Denote by $\mathrm{Tri}(X)$ the space of triples of distinct points in a space X. The action $G \curvearrowright \partial_\infty G$ induces an action $G \curvearrowright \mathrm{Tri}(\partial_\infty G)$. This action is discrete and cocompact, and Gromov hyperbolic groups are characterized by this property according to a theorem by Bowditch [Bow].

5. Cannon's conjecture and fractal 2-spheres

It is a natural question to what extent the structure of a Gromov hyperbolic group G is reflected in its boundary $\partial_\infty G$. For example, $\partial_\infty G$ is totally disconnected iff G is virtually free, i.e., it contains a free group of finite index. Similarly, $\partial_\infty G$ is a topological circle iff G is virtually Fuchsian (see [KB] for these and related results). The case where $\partial_\infty G$ is a topological 2-sphere is covered by the following conjecture [Ca].

Conjecture 5.1 (*Cannon's conjecture, Version* I). Suppose G is a Gromov hyperbolic group whose boundary at infinity $\partial_\infty G$ is homeomorphic to $\mathbb{S}^2$. Then there exists an action of G on hyperbolic 3-space $\mathbb{H}^3$ that is isometric, properly discontinuous, and cocompact.

If true, this conjecture would essentially give a characterization of fundamental groups of closed hyperbolic 3-orbifolds from the point of view of geometric group theory. The conjecture is equivalent to a quasisymmetric uniformization problem.

Conjecture 5.2 (*Cannon's conjecture, Version* II). Suppose G is a Gromov hyperbolic group whose boundary at infinity $\partial_\infty G$ is homeomorphic to $\mathbb{S}^2$. Then $\partial_\infty G$ is quasisymmetrically equivalent to $\mathbb{S}^2$.

Indeed, if Conjecture 5.2 holds, then we can conjugate the natural action of G on $\partial_\infty G$ to a uniformly quasiconformal action of G on $\mathbb{S}^2$. By a well-known theorem due

to Sullivan [Su1] and to Tukia [Tu] such an action is conjugate to an action of G on $\mathbb{S}^2$ by Möbius transformations. Considering $\mathbb{S}^2$ as the boundary at infinity of $\mathbb{H}^3$, we can extend this action to an isometric action of G on $\mathbb{H}^3$ with the desired properties.

Conversely, if G acts on $\mathbb{H}^3$ isometrically, properly discontinuously, and cocompactly, then the Cayley graph of G with respect to any (finite and symmetric) set of generators is quasi-isometric to $\mathbb{H}^3$. This quasi-isometry induces the desired quasisymmetric equivalence between $\partial_\infty G$ and $\partial_\infty \mathbb{H}^3 = \mathbb{S}^2$.

Cannon, Floyd, and Parry [CFP] have attempted to settle Conjecture 5.1 by using subdivision rules and Cannon's Combinatorial Riemann Mapping Theorem [Ca]. A different approach is due to B. Kleiner and the author. In [BK4] recently developed techniques from analysis on metric spaces were used and led to the following theorem.

Theorem 5.3. *Suppose G is a Gromov hyperbolic group whose boundary at infinity $\partial_\infty G$ is homeomorphic to $\mathbb{S}^2$. If the Ahlfors regular conformal dimension of $\partial_\infty G$ is attained as a minimum, then ∂G_∞ is quasisymmetrically equivalent to $\mathbb{S}^2$.*

Here the (*Ahlfors regular*) *conformal dimension* of a metric space X is defined as

$$\dim_{\mathrm{c}} X = \inf\{Q : \text{there exists an Ahlfors } Q\text{-regular metric space } Y \text{ that is quasisymmetrically equivalent to } X\}. \tag{5}$$

This concept was implicitly introduced by Bourdon and Pajot [BP]. Pansu [Pa2] has defined a related, but different concept of conformal dimension of a space. If $X = \partial_\infty G$, where G is a Gromov hyperbolic group, then the set over which the infimum in (5) is taken is nonempty, because the boundary of a Gromov hyperbolic group equipped with a visual metric is Ahlfors regular [Co].

In more intuitive terms, the above result can be formulated as follows: Let G be a Gromov hyperbolic group as in Cannon's conjecture. If a certain infimum for $\partial_\infty G$ is attained as a minimum (related to how much we can "squeeze" the space by a quasisymmetric map while retaining Ahlfors regularity for some exponent), then the desired conclusion holds.

The proof of this theorem depends on a recent result by Keith and Laakso [KL] which essentially says that if $Q > 1$ is the conformal dimension of a Q-regular space X, then X has a *weak tangent* (see [BBI, Ch. 8] for the definition and [BK2] for related discussions), carrying a family of non-constant paths with positive Q-modulus.

In our situation $X = \partial_\infty G$ is the boundary of a Gromov hyperbolic group G, and is equipped with a metric that comes from the minimizer in (5). Every weak tangent of X is quasisymmetrically equivalent to $\partial_\infty G$ minus a point [BK2]. Therefore, $\partial_\infty G$ itself carries a family of positive Q-modulus, where Q is the conformal dimension of $\partial_\infty G$. The main work now consists in showing that the natural group action $G \curvearrowright \partial_\infty G$ allows one to promote this to the stronger conclusion that $\partial_\infty G$ has families of non-constant paths with uniformly positive Q-modulus on all locations and scales; more precisely, that $X = \partial_\infty G$ is a Q-regular Q-Loewner space. Up to this point, the assumption that $\partial_\infty G$ is homeomorphic to $\mathbb{S}^2$ was not used. The proof

of the above statement is now finished by invoking Theorem 3.3 which shows that $\partial_\infty G$ is quasisymmetrically equivalent to $\mathbb{S}^2$.

Related to these questions is the following problem due to P. Papasoglu: Suppose that G is a Gromov hyperbolic group whose boundary $\partial_\infty G$ is homeomorphic to $\mathbb{S}^2$. Cannon's conjecture predicts that in this situation $\partial_\infty G$ is quasisymmetrically equivalent to $\mathbb{S}^2$; in particular, $\partial_\infty G$ should contain many quasicircles. While Cannon's conjecture is still open, can one at least prove that $\partial_\infty G$ contains a *single* quasicircle? The following result proved in [BK5] settles this in the affirmative.

Theorem 5.4. *The boundary of a Gromov hyperbolic group contains a quasicircle if and only if the group is not virtually free.*

In order to get a better understanding of the relevant issues in Cannon's conjecture, it seems natural to study uniformly quasi-Möbius actions on compact metric spaces X such that the induced action on the space $\mathrm{Tri}(X)$ of triples is discrete and cocompact. In addition to these assumptions it is reasonable to require that X is Ahlfors regular.

If a metric space X is Q-regular, then the exponent Q is at least as big as the topological dimension of X. The borderline case where Q equals the topological dimension of X is of particular interest. In [BK2] (for $n \geq 2$) and [BK3] (for $n = 1$) the following rigidity theorem was proved in all dimensions (see [Su2] and [Yu] for related results).

Theorem 5.5. *Let X be a compact, Ahlfors n-regular metric space of topological dimension $n \in \mathbb{N}$. Suppose that a group G acts on X by uniformly quasi-Möbius maps and that the induced action on* $\mathrm{Tri}(X)$ *is discrete and cocompact. Then the action $G \curvearrowright X$ is quasisymmetrically conjugate to a Möbius group action on the standard sphere $\mathbb{S}^n$.*

Note that we do not assume that X is homeomorphic to $\mathbb{S}^n$. We get the quasisymmetric equivalence of X and $\mathbb{S}^n$ as part of the conclusion.

6. Post-critically finite rational maps

Apart from Gromov hyperbolic groups, there are other dynamical systems where quasisymmetric uniformization problems arise. Interesting examples are provided by post-critically finite rational maps R on the Riemann sphere $\overline{\mathbb{C}}$ [DH].

Suppose $R\colon \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a holomorphic map of $\overline{\mathbb{C}}$ onto itself, i.e., a rational function. Let Ω_R denote the set of critical points of R, and $P_R = \bigcup_{n\in\mathbb{N}} R^n(\Omega_R)$ be the set of post-critical points of R (here R^n denotes the n-th iterate of R). We make the following assumptions on R:

(i) R is post-critically finite, i.e., P_R is a finite set,

(ii) R has no periodic critical points; this implies that $\mathcal{J}(R) = \overline{\mathbb{C}}$ for the Julia set of R,

(iii) the orbifold $\mathcal{O}_R$ associated with R is hyperbolic (see [DH] for the definition of $\mathcal{O}_R$); this implies that the dynamics of R on the Julia set $\mathcal{J}(R) = \overline{\mathbb{C}}$ is expanding.

A characterization of post-critically finite rational maps is due to Thurston. The right framework is the theory of topologically holomorphic self-maps $f\colon \mathbb{S}^2 \to \mathbb{S}^2$ of the sphere. By definition these maps have the local form $z \mapsto z^n$ with $n \in \mathbb{N}$ in appropriate local coordinates, and one defines the critical set, the post-critical set, and the associated orbifold similarly as for rational maps. In our context, Thurston's theorem can be stated as follows [DH].

Theorem 6.1. *Let $f\colon \mathbb{S}^2 \to \mathbb{S}^2$ be a post-critically finite topologically holomorphic map with hyperbolic orbifold. Then f is equivalent to a rational map R if and only if f has no "Thurston obstructions".*

Equivalence has to be understood in an appropriate sense. If f and R are both expanding, this just means conjugacy of the maps.

A Thurston obstruction is defined as follows. A *multicurve* $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ is a system of Jordan curves in $\mathbb{S}^2 \setminus P_f$ with the following properties: the curves have pairwise empty intersection, are pairwise non-homotopic in $\mathbb{S}^2 \setminus P_f$, and non-peripheral (this means that each of the complementary components of a curve contains at least two points in P_f). A multicurve Γ is called *f-stable* if for all j every component of $f^{-1}(\gamma_j)$ is either peripheral or homotopic in $\mathbb{S}^2 \setminus P_f$ to one of the curves γ_i.

If Γ is an f-stable multicurve, fix i and j and label by α the components $\gamma_{i,j,\alpha}$ of $f^{-1}(\gamma_j)$ homotopic to γ_i in $\mathbb{S}^2 \setminus P_f$. Then f restricted to $\gamma_{i,j,\alpha}$ has a mapping degree $d_{i,j,\alpha} \in \mathbb{N}$. Let

$$m_{ij} = \sum_{\alpha} \frac{1}{d_{i,j,\alpha}}$$

and define the *Thurston matrix* $A(\Gamma)$ of the f-stable multicurve Γ by $A(\Gamma) = (m_{ij})$. This is a matrix with nonnegative coefficients; therefore, it has a largest eigenvalue $\lambda(f, \Gamma) \geq 0$. Then Γ is a *Thurston obstruction* if $\lambda(f, \Gamma) \geq 1$.

In Figure 1 we see topological 2-spheres obtained by gluing together 16 squares (colored black and white in an alternating fashion) for the surface on the left, and two large squares for the surface on the right. The map f is constructed by scaling a white square so that it corresponds to the top square on the right and extending the partially defined map to the whole surface by "Schwarz reflection". Then f is post-critically finite with a set of 4 post-critical points (the corners of the large squares). An f-stable multicurve Γ consisting of one Jordan curve γ is indicated on the right. It has 4 preimages on the left. Two of them are peripheral, and the other ones are homotopic to γ in the complement of P_f. Since the degree of the map on these curves is 2, the Thurston matrix is a (1×1)-matrix with the entry $1/2 + 1/2 = 1$. Hence Γ is a Thurston obstruction and f is not equivalent to a rational map.

Post-critically finite rational maps are related to *subdivision rules* [CFKP], [CFP]. For example, if R is a real rational map, i.e., $R(\overline{\mathbb{R}}) \subseteq \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, satisfying

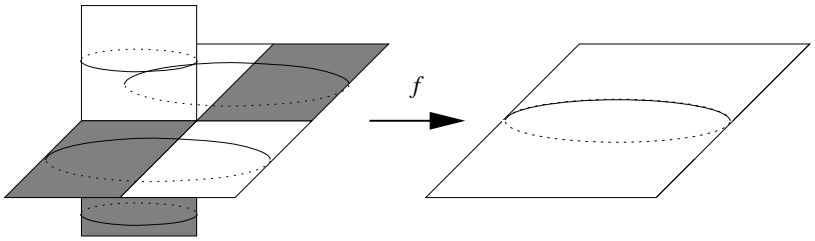


Figure 1. A post-critically finite topologically holomorphic map with a Thurston obstruction.

the above conditions (i)–(iii), and $P_R \subseteq \overline{\mathbb{R}}$, then $R^{-1}(\overline{\mathbb{R}})$ is a graph providing a subdivision of the upper and lower half-planes. It will give rise to a *one-tile subdivision rule*, because the upper and the lower half-planes are subdivided in the same way. The combinatorics of the graphs $R^{-n}(\overline{\mathbb{R}})$ with its corresponding tiles of level n (the closures of the complementary components of the graph $R^{-n}(\overline{\mathbb{R}})$) is determined by iterating the subdivision rule n times. Note that once the subdivision rule is given, the map R admits a completely combinatorial description as an "expanding map" of the subdivision rule by specifying which tiles on level n are mapped to which tiles on level $n-1$. The map f in Figure 1 is also associated with a one-tile subdivision rule which describes how the squares on the right are subdivided into 8 squares each to obtain the combinatorics of the squares on the surface on the left.

For general, not necessarily real rational functions, one expects at least two tile types. More precisely, one can ask whether every rational map satisfying (i)–(iii), or at least a sufficiently high iterate of such a map, is associated with a two-tile subdivision rule. This is indeed the case [BMy], showing that the behavior of the rational maps as discussed admits a combinatorial description.

Theorem 6.2. *Let R be a rational function satisfying* (i)–(iii). *Then there exists an iterate R^n and a quasicircle $C \subseteq \overline{\mathbb{C}}$ such that $P_{R^n} \subseteq C$ and $R^n(C) \subseteq C$.*

A related result has been announced by Cannon, Floyd, and Parry (unpublished).

Conversely, one can start with a two-tile subdivision rule of $\mathbb{S}^2$ (satisfying additional technical assumptions encoding the properties (i)–(iii)). One can associate a natural metric d_λ on $\mathbb{S}^2$ with such a subdivision rule. Roughly speaking, one fixes a parameter $\lambda < 1$ and declares tiles on level n to have size λ^n. The distance $d_\lambda(x, y)$ between two points $x, y \in \mathbb{S}^2$ is then defined as the infimum of all sums of tile-sizes in chains of tiles connecting x and y. Here one has to allow tiles of different levels in a chain. If $\lambda < 1$ is sufficiently close to 1, this gives a metric d_λ on $\mathbb{S}^2$ such that the diameter of a tile on level n is comparable to λ^n. The ambiguity in the parameter λ is not very serious and leads to quasisymmetrically equivalent metrics. These metrics d_λ form an analog of the visual metrics on the boundary of a Gromov hyperbolic group.

If we denote by X the sphere $\mathbb{S}^2$ equipped with this metric, then X is Ahlfors regular and linearly locally connected, and the subdivision rule produces a topologically holomorphic expanding map $f\colon X \to X$ which is post-critically finite, and which is "uniformly" quasiregular with respect to a suitable notion of quasiregularity in this metric space context. The question when the dynamical system $f\colon X \to X$ comes from a rational map can be formulated as a quasisymmetric uniformization problem. The following result is essentially contained in [My].

Theorem 6.3. *The map $f\colon X \to X$ is conjugate to a rational map $R\colon \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ if and only if X is quasisymmetrically equivalent to $\mathbb{S}^2$.*

If f is not equivalent to a rational function, then it is natural to ask whether the dynamical system $f\colon X \to X$ is conjugate to a corresponding dynamical system on a better and less "fractal" space. More precisely, we want to replace X by a quasisymmetrically equivalent Ahlfors regular space of lower Hausdorff dimension. As in Theorem 5.3 discussed above, this leads to the problem of finding the conformal dimension $\dim_{\mathrm{c}} X$ of the self-similar space X. By Theorem 6.3 the conformal case is characterized by the fact that we can squeeze X to a 2-regular space (and hence to the standard sphere $\mathbb{S}^2$ according to Theorem 3.2) by a quasisymmetric map. So we have a situation that is very similar to Cannon's conjecture.

In discussions with L. Geyer and K. Pilgrim the following conjecture for $\dim_{\mathrm{c}} X$ in terms of dynamical data emerged. To state it, let $Q \geq 2$ and Γ be an f-stable multicurve, define the modified Thurston matrix $A(\Gamma, Q)$ as $A(\Gamma, Q) = (m_{ij}^Q)$, where

$$m_{ij}^Q = \sum_\alpha \frac{1}{d_{i,j,\alpha}^{Q-1}},$$

and let $\lambda(f, \Gamma, Q)$ be the largest nonnegative eigenvalue of $A(\Gamma, Q)$.

Conjecture 6.4. If X comes from a subdivision rule with associated expanding map f, then $\dim_{\mathrm{c}} X$ is the infimum of all $Q \geq 2$ such that $\lambda(f, \Gamma, Q) < 1$ for all f-stable multicurves Γ.

As in the proof of Theorem 6.1 (related to the necessity of the condition), there is one part of Conjecture 6.4 that seems to be rather easy to establish: If there exists an f-stable multicurve Γ with $\lambda(f, \Gamma, Q) \geq 1$, then $\dim_{\mathrm{c}} X \geq Q$. The idea for proving this is to find path families related to ring domains associated with Γ which have positive Q-modulus. Any such path family on an Ahlfors regular space provides an obstruction to lowering its dimension by a quasisymmetric map [He, Thm. 15.10].

For some nontrivial cases one can show that Conjecture 6.4 is true [BMy]. For example, if X is the fractal obtained from the subdivision rule suggested by Figure 1, then $\dim_{\mathrm{c}} X = 2$. This corresponds to the prediction of Conjecture 6.4 in this case. Note that here the infimum defining $\dim_{\mathrm{c}} X$ is not attained as a minimum; otherwise f would be equivalent to a rational map by Theorem 6.3. We have seen above that this is not the case.

7. Sierpiński carpets

Sierpiński carpets are fractal spaces with a very interesting quasiconformal geometry. Recall that the "standard" Sierpiński carpet is obtained as follows: Start with the closed unit square, and subdivide it into $9 = 3 \times 3$ equal subsquares. Remove the interior of the middle square, and repeat this procedure for each of the remaining 8 subsquares. The limiting object of this construction is the standard Sierpiński carpet.

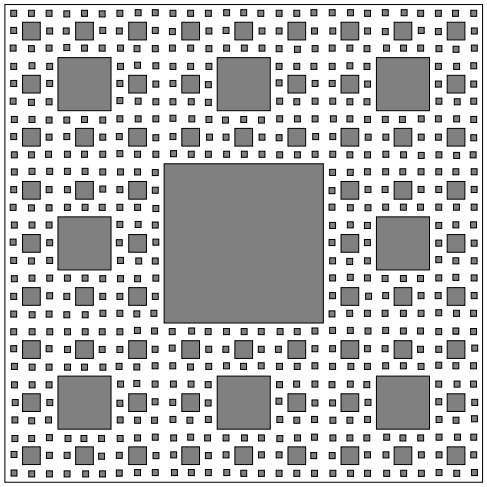


Figure 2. The standard Sierpiński carpet S_3.

One can run a similar construction, where one subdivides each square into $(p \times p)$-subsquares, p odd, and removes the middle square in each step. We denote the resulting space by S_p. So the standard Sierpiński carpet is S_3. We equip S_p with the restriction of the Euclidean metric on $\mathbb{R}^2$.

The spaces S_p are all homeomorphic to each other as follows from the following topological characterization theorems due to Whyburn [Wh].

Theorem 7.1. *Let X be a metric space. Then X is homeomorphic to the standard Sierpiński carpet if and only if X is a locally connected continuum, is topologically planar, has topological dimension* 1*, and has no local cut points.*

Here we call a set *topologically planar* if it is homeomorphic to a subset of the plane $\mathbb{R}^2$. A *local cut point* p of X is a point that has a connected neighborhood U such that $U \setminus \{p\}$ is not connected.

Theorem 7.2. *Let $X = \mathbb{S}^2 \setminus \bigcup_{i\in\mathbb{N}} D_i$ be the complement in $\mathbb{S}^2$ of countably many pairwise disjoint open Jordan regions D_i. Then X is homeomorphic to the standard Sierpiński carpet if and only if X has empty interior, $\partial D_i \cap \partial D_j = \emptyset$ for $i \neq j$, and* $\operatorname{diam}(D_i) \to 0$ *as* $i \to \infty$.

In the following we will call a metric space X a *carpet* if it is homeomorphic to S_3. A topological circle J in a carpet X is called a *peripheral circle* if J does not separate X, i.e., if $X \setminus J$ is connected. If X is a carpet as in Theorem 7.2, then the peripheral circles of X are precisely the Jordan curves ∂D_i, $i \in \mathbb{N}$. Note that every homeomorphism between two carpets X and Y has to map every peripheral circle of X to a peripheral circle of Y.

The deeper reason why all spaces as in Theorem 7.1 are homeomorphic to each other is that the homeomorphisms on a carpet form a rather large and flexible class. This is illustrated by the following transitivity result: If X is a carpet, $\{C_1, \dots, C_n\}$ and $\{C_1', \dots, C_n'\}$ are two collections of peripheral circles of X with n elements, then there exists a homeomorphism $f\colon X \to X$ such that $f(C_i) = C_i'$. In other words, the homeomorphism group of X acts n-transitively on the set of peripheral circles of X for every $n \in \mathbb{N}$. This changes drastically if one restricts attention to quasisymmetric homeomorphisms and surprising rigidity phenomena emerge (cf. the Three-Circle Theorem 8.3 below). To discuss instances of this we first introduce some more terminology.

We say that a carpet $X \subseteq \mathbb{S}^2$ is *round* if its peripheral circles are round circles, i.e., if X is as in Theorem 7.2, where the Jordan regions D_i are round disks. If X is a round carpet and $f\colon \mathbb{S}^2 \to \mathbb{S}^2$ is a (possibly orientation reversing) Möbius transformation, then $f(X)$ is also a round carpet. We say that a round carpet X is *rigid* if this is the only way to obtain another round carpet as a quasisymmetric image of X, i.e., if every quasisymmetric map $g\colon X \to Y$ to another round carpet Y is the restriction of a Möbius transformation. Rigid round carpets admit a simple characterization [BKM].

Theorem 7.3. *A round carpet X is rigid if and only if it has measure zero.*

Actually, in [BKM] a related rigidity result is proved in all dimensions. Call a subset $X \subseteq \mathbb{S}^n$, $n \geq 2$, a *Schottky set* if it is the complement of pairwise disjoint open balls, and call a Schottky set $X \subseteq \mathbb{S}^n$ *rigid* if every quasisymmetric map $f\colon X \to Y$ to another Schottky set $Y \subseteq \mathbb{S}^n$ is the restriction of a Möbius transformation. Then one can show that every Schottky set of measure zero is rigid. This is a strengthening of a result due M. Kapovich, B. Kleiner, B. Leeb, and R. Schwartz (unpublished).

A corollary of Theorem 7.3 is that the set of quasisymmetric equivalence classes of round carpets has the cardinality of the continuum. So even though topologically there is only one Sierpiński carpet, from the point of view of quasiconformal geometry, there are many different ones.

An important source of round carpets is the theory of Kleinian groups. Let M be a compact hyperbolic 3-orbifold with nonempty totally geodesic boundary. Its universal cover $\tilde{M}$ is isometric to a convex subset K of $\mathbb{H}^3$ bounded by a nonempty collection of pairwise disjoint hyperplanes. Then the boundary at infinity $\partial_\infty K \subseteq \partial_\infty \mathbb{H}^3 = \mathbb{S}^2$ of K is a round carpet. The fundamental group $G = \pi_1(M)$ of M acts in a natural way on K by isometries. This induces an action $G \curvearrowright \partial_\infty K$ of G on the round carpet $S = \partial_\infty K$ by Möbius transformations. The group G is Gromov hyperbolic and its boundary $\partial_\infty G$ is quasisymmetrically equivalent to S. Hence the group $\mathrm{QS}(S) \supseteq G$

of quasisymmetric self-maps of S is rather large, because it acts cocompactly on triples of S and so there are only finitely many distinct orbits of peripheral circles. Accordingly, one should think of these round carpets as particularly "symmetric" ones.

It is tempting to try to characterize this situation from the point of view of Gromov hyperbolic groups. An analog of Cannon's conjecture is the following conjecture due to Kapovich and Kleiner who studied Gromov hyperbolic groups with carpet boundaries [KK]: *Suppose G is a Gromov hyperbolic group such that $\partial_\infty G$ is a carpet. Then G admits a properly discontinuous, cocompact and isometric action on a convex subset of $\mathbb{H}^3$ with nonempty totally geodesic boundary.*

This can be reformulated as a quasisymmetric uniformization problem.

Conjecture 7.4 (*Kapovich–Kleiner conjecture).* Suppose G is a Gromov hyperbolic group with $\partial_\infty G$ homeomorphic to the standard Sierpiński carpet. Then $\partial_\infty G$ is quasisymmetrically equivalent to a round carpet.

We call a carpet a *group carpet* if it arises as (i.e., is quasisymmetrically equivalent to) a boundary of a Gromov hyperbolic group. So the Kapovich–Kleiner conjecture asks whether every group carpet is quasisymmetrically equivalent to a round carpet.

Group carpets X should be thought of as very self-similar fractal spaces. As in the Kleinian case, the group $\mathrm{QS}(X)$ of quasisymmetric self-maps of X is rather large. It acts cocompactly on triples, and so there are only finitely many distinct orbits of peripheral circles. In addition, the collection of peripheral circles of a group carpet has the following geometric properties:

(i) The peripheral circles are *uniform quasicircles*, i.e., each one is quasisymmetrically equivalent to $\mathbb{S}^1$ by an η-quasisymmetric map with η independent of the peripheral circle.

(ii) The peripheral circles are *uniformly separated*, i.e., there is a uniform positive lower bound for the relative distance

$$\frac{\mathrm{dist}(C, C')}{\min\{\mathrm{diam}(C), \mathrm{diam}(C)\}}$$

of two distinct peripheral circles C and C'.

(iii) The peripheral circles *occur on all locations and scales*, i.e., if B is a ball in the carpet, then there exists a peripheral circle that intersects B and has a size comparable to B.

In view of the Kapovich–Kleiner conjecture one can ask whether these conditions are sufficient for X to be quasisymmetrically equivalent to a round carpet. It turns out that this is true for carpets that can be quasisymmetrically embedded into $\mathbb{S}^2$. This is a consequence of the following uniformization result [Bo].

Theorem 7.5. *Let $X \subseteq \mathbb{S}^2$ be a carpet, and suppose that the peripheral circles of X are uniform quasicircles and are uniformly separated. Then there exists a quasisymmetric map $f\colon X \to Y$ to a round carpet $Y \subseteq \mathbb{S}^2$.*

This theorem applies for example to the carpets S_p. So they are quasisymmetrically equivalent to round carpets. Note that if X as in the theorem has measure zero in addition (which is true if X is quasisymmetrically equivalent to a group carpet), then the uniformizing map f is uniquely determined up to a post-composition by a Möbius transformation (this essentially follows from Theorem 7.3). This shows that one can expect very little flexibility in constructing the uniformizing map f.

Theorem 7.5 is an analog of Koebe's well-known result on uniformization by circle domains. It says that every region in $\mathbb{S}^2$ with finitely many complementary components is conformally equivalent to a *circle domain*, i.e., a region whose complementary components are round (possibly degenerate) disks. This statement is actually used in the proof of Theorem 7.5. One considers regions Ω_n obtained by removing from $\mathbb{S}^2$ the closures of n complementary components of the given carpet X. By circle uniformization one can map the regions Ω_n to circle domains by (suitably normalized) conformal maps f_n. The uniformizing map f of X to a round carpet is then obtained as a sublimit of the sequence of maps f_n. The main difficulty is to show that such a sublimit exists. For this one proves that the maps f_n are uniformly quasisymmetric, i.e., η-quasisymmetric with η independent of n. It is a standard idea to use modulus estimates to get the required uniform distortion estimates for the maps f_n. If X has measure zero, then X does not support path families of positive modulus. Accordingly, one cannot expect any control for the distortion coming from such estimates involving classical modulus. This situation is remedied by a new quasisymmetric invariant, the modulus of a path family with respect to a carpet, which is the main technical ingredient in the proof of Theorem 7.5.

Let $X \subseteq \mathbb{S}^2$ be a carpet with peripheral circles C_i, $i \in \mathbb{N}$, and Γ a family of paths in $\mathbb{S}^2$. Then the *modulus of* Γ *with respect to* X is defined as

$$M_X(\Gamma) = \inf\Big\{ \sum_{i\in\mathbb{N}} \rho_i^2 : \rho = \{\rho_i\} \text{ is admissible for } \Gamma \Big\}.$$

Here a sequence $\rho = \{\rho_i\}$ of nonnegative weights ρ_i is called *admissible for* Γ if there exists an exceptional path family $\Gamma_0 \subseteq \Gamma$ with $\operatorname{Mod}_2(\Gamma_0) = 0$ such that

$$\sum_{\gamma\cap C_i\neq\emptyset} \rho_i \geq 1$$

for all paths $\gamma \in \Gamma \setminus \Gamma_0$.

So in contrast to classical modulus where ρ is a density, the test function is an assignment of discrete weights ρ_i to the peripheral circles C_i. This is similar to Schramm's notion of "transboundary extremal length" [Sc], where the test function consists both of a density and a discrete part. In the definition of $M_X(\Gamma)$ one wants

to infimize the total mass $\sum_{i\in\mathbb{N}} \rho_i^2$ for all admissible sequences $\rho = \{\rho_i\}$. The admissibility requires that essentially every path picks up at least total weight 1 from all the peripheral circles that it meets. An important subtlety here is to allow the exceptional path family Γ_0. Otherwise, the quantity $M_X(\Gamma)$ would be infinite (and hence useless) for sufficiently large families Γ.

In contrast to classical modulus which is distorted by a multiplicative amount (cf. Theorem 2.1), the quantity $M_X(\Gamma)$ is invariant under quasisymmetric maps on $\mathbb{S}^2$.

Proposition 7.6. *Let $X \subseteq \mathbb{S}^2$ be a carpet, Γ a path family in $\mathbb{S}^2$, and $f : \mathbb{S}^2 \to \mathbb{S}^2$ a quasisymmetric map. Then*

$$M_X(\Gamma) = M_{f(X)}(f(\Gamma)).$$

The restriction to global maps f is not very serious here if one requires that the peripheral circles of X are uniform quasicircles. Then one can extend every quasisymmetric embedding of X into $\mathbb{S}^2$ to a quasisymmetric homeomorphism on $\mathbb{S}^2$.

As we remarked, Theorem 7.5 would settle the Kapovich–Kleiner conjecture if one could always quasisymmetrically embed a group carpet into $\mathbb{S}^2$. The conditions (i)–(iii) for the peripheral circles discussed above are not enough to guarantee this, because there are some carpets with these properties which do not admit such an embedding. In the positive direction one can show that if $\dim_{\mathrm{c}} X < 2$ for such a carpet X, then one gets the desired quasisymmetric embedding into $\mathbb{S}^2$. This was recently proved by B. Kleiner and the author [BK6]. The idea for the proof (due to J. Heinonen) is that each peripheral circle can be filled by a metric disk to obtain a sphere to which Theorem 3.2 can be applied. To get fillings of the right type one uses conformal densities as in [BHR].

8. Rigidity of square carpets

A carpet $X \subseteq \mathbb{R}^2$ is called a *square carpet* if its peripheral circles are boundaries of squares. Examples are the carpets $S_3, S_5, \ldots$ introduced in the previous section. Obviously, these carpets are very symmetric and self-similar, so one may wonder whether they are group carpets. If so, the groups $\mathrm{QS}(S_p)$ should be rather large, and in particular infinite. It turns out that this is not the case [BMe].

Theorem 8.1. *Suppose $f : S_3 \to S_3$ is a quasisymmetric map. Then f is an isometry.*

The only isometries of S_3 are the obvious symmetries given by reflections and rotations; so $\mathrm{QS}(S_3)$ is a dihedral group containing 8 elements. It is very likely that an analog of Theorem 8.1 is true for all carpets S_p, p odd. At the moment it is only known that $\mathrm{QS}(S_p)$ is always a finite dihedral group. This implies that no carpet S_p is a group carpet.

The proof of Theorem 8.1 is surprisingly difficult. To explain some of the ingredients, suppose $f : S_3 \to S_3$ is a quasisymmetric map. For simplicity assume that f

is orientation preserving. Denote by C_1 the boundary of the unit square, and by C_2 the boundary of the middle square that was deleted in the first step of the construction of S_3. So C_1 and C_2 are peripheral circles of S_3.

If C and C' are two distinct peripheral circles of S_3, let $\Gamma(C, C')$ be the family of all open paths $\gamma|_{(0,1)}$, where $\gamma\colon [0, 1] \to \mathbb{S}^2$ is a path connecting C and C' such that $\gamma(0) \in C$, $\gamma(1) \in C'$, and $\gamma(0, 1) \cap (C \cup C') = \emptyset$. Then the (unordered) pair $\{C_1, C_2\}$ is distinguished from all other pairs $\{C, C'\}$ due to the following fact.

Lemma 8.2. *If C and C' are two distinct peripheral circles of S_3, then*

$$M_{S_3}(\Gamma(C, C')) \le M_{S_3}(\Gamma(C_1, C_2))$$

with equality if and only if $\{C, C'\} = \{C_1, C_2\}$.

The proof crucially uses the self-similarity of S_3 combined with monotonicity properties of the modulus invariant $M_X(\Gamma)$ defined in the previous section.

An immediate consequence of Lemma 8.2 and Proposition 7.6 is that

$$\{f(C_1), f(C_2)\} = \{C_1, C_2\}.$$

In other words, f preserves the peripheral circles C_1 and C_2 setwise or exchanges them.

Let us again make a simplifying assumption, namely that $f(C_1) = C_1$ and $f(C_2) = C_2$. Now one analyzes the possibilities for the images of the eight peripheral circles of S_3 that constitute the boundaries of the squares deleted in the second step of the construction of S_3. These eight peripheral circles come in two groups: "corner" circles and "side" circles. Using ideas as in the proof of Lemma 8.2, one can show that at least one of these eight second-generation peripheral circles is mapped to another second-generation peripheral circle; say one of the corner circles C_3 is mapped to a corner circle or a side circle C_3'.

In the first case where C_3' is also a corner circle there exists a rotation R of S_3 such that $R(C_3) = C_3'$. Since R also preserves C_1 and C_2 setwise, we conclude that $f = R$ by the following theorem (applied to $g = R^{-1} \circ f$).

Theorem 8.3 (Three-circle theorem). *Let $X \subseteq \mathbb{S}^2$ be a carpet of measure zero whose peripheral circles are uniform quasicircles and are uniformly separated. Suppose C_1, C_2, C_3 are three distinct peripheral circles of X and $g\colon X \to X$ is an orientation preserving quasisymmetric map such that $g(C_i) = C_i$, $i = 1, 2, 3$. Then g is the identity on X.*

In other words, if an orientation preserving quasisymmetric map of the carpet fixes three peripheral circles setwise, then it is the identity. The same proof will show that if g fixes three points (instead of three peripheral circles), then the same conclusion holds, i.e., g is the identity.

Proof. By Theorem 7.5 there exists a quasisymmetric uniformization map $h\colon X \to Y$ to a round carpet $Y \subseteq \mathbb{S}^2$. One can show that the map h can be extended to a global quasiconformal map $H\colon \mathbb{S}^2 \to \mathbb{S}^2$. Since quasiconformal maps preserve sets of measure zero, the round carpet Y has measure zero. Hence Y is rigid by Theorem 7.3. Therefore, the quasisymmetric map $\tilde{g} = h \circ g \circ h^{-1}\colon Y \to Y$ is the restriction of an orientation-preserving Möbius transformation. Since it fixes the three round circles $h(C_i)$ setwise, it is the identity on Y. Hence g is the identity on X. □

The second case where the corner circle C_3 is mapped to a side circle C_3' does not occur. To rule out the existence of such a "ghost" map, one argues by contradiction. Suppose the situation is as represented in Figure 3. If R_D and R_M denote the reflections in the indicated symmetry lines D and M of S_3, respectively, one can show that

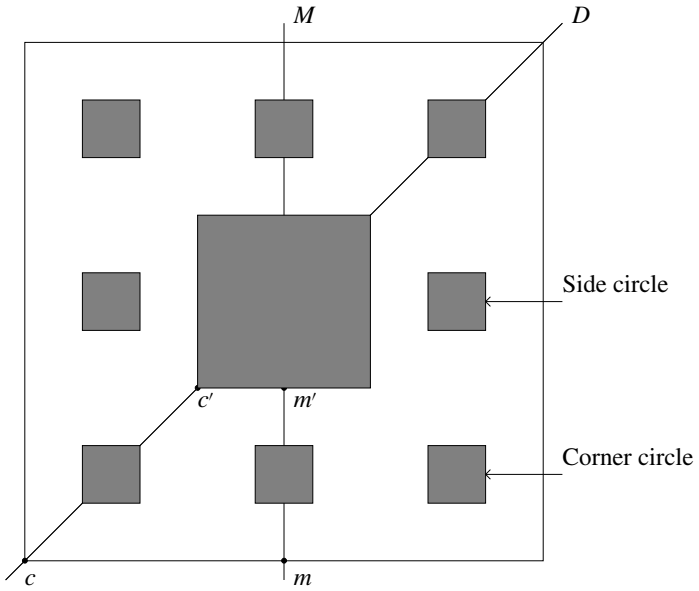


Figure 3. Corner and side circles of S_3.

$R_M \circ f = f \circ R_D$ by using the Three-Circle Theorem 8.3. This leads to $f(c) = m$ and $f(c') = m'$. Blowing up the carpet at c and m, the map f induces a quasisymmetric equivalence between the weak tangents W_c and W_m of S_3 at these points (with a suitable normalization these weak tangents are uniquely determined up to isometry). Similarly, the weak tangents $W_{c'}$ and $W_{m'}$ are quasisymmetrically equivalent. Since W_m and $W_{m'}$ are isometric, one concludes that W_c and $W_{c'}$ are quasisymmetrically equivalent. Since $W_{c'}$ essentially consists of three copies of W_c, one can get the desired contradiction by using the modulus invariant $M_X(\Gamma)$ and its monotonicity properties. The last step in the proof could be simplified, if one knew that W_c and W_m

are not quasisymmetrically equivalent. This is likely to be true, but an open problem at the moment.

The following result was also proved in [BMe].

Theorem 8.4. *Let $p, q \geq 3$ be odd integers. Then S_p and S_q are quasisymmetrically equivalent if and only if $p = q$.*

Using the known estimate

$$\dim_{\mathrm{c}} S_p \geq 1 + \frac{\log(p-1)}{\log p}$$

for the conformal dimension of S_p, it is not hard to see that S_p cannot be quasisymmetrically equivalent to S_q if p is much larger than q. The full result Theorem 8.4 is much harder to establish and uses ideas similar to the ones just described. An interesting open problem in this connection is to determine $\dim_{\mathrm{c}} S_p$.

9. Conclusion

It is evident from the preceding discussion that we are still far from a full understanding of the quasiconformal geometry of fractal 2-spheres and Sierpiński carpets. An obstacle in the solution of Cannon's conjecture is the lack of examples that could reveal some hidden structures. All known examples of Gromov hyperbolic groups with 2-sphere boundary arise from the standard Kleinian group situation, and Cannon's conjecture predicts that there are no others. In this sense the fractal 2-spheres that arise in the dynamics of post-critically finite maps exhibit more interesting phenomena, because sometimes they are quasisymmetrically equivalent to the standard 2-sphere and sometimes not. By investigating these spaces one may discover some general obstruction (formed by a "large" path family for example) that prevents a self-similar 2-sphere from obtaining a minimum for its conformal dimension. One may speculate that in the situation of Cannon's conjecture the group action prevents the existence of such an obstruction. This would lead to a solution of the conjecture according to Theorem 5.3.

The Kapovich–Kleiner conjecture looks somewhat more accessible due to the additional features given by the geometry of the peripheral circles of a group carpet. It has to be explored whether a modulus invariant similar to the invariant $M_X(\Gamma)$ for carpets in the plane can be used to prove general uniformization theorems for metric carpets.

The picture is sketchiest for the rigidity results on square carpets. Here it would be desirable to put isolated facts such as Theorems 8.1 and 8.4 into a general framework. A possible venue here is to develop an analytic theory of quasisymmetrically invariant "harmonic" functions. Similarly as in the definition of $M_X(\Gamma)$ this can be based on the minimization of energies of discrete weights $\rho = \{\rho_i\}$ which play the role of

"upper" gradients of the functions. Such a theory could also lead to the solution of problems about weak tangents of carpets such as the one mentioned in Section 8.

References

[Ah] Ahlfors, L. V., *Collected papers*. Volume I, Contemp. Mathematicians, Birkhäuser, Boston 1982.

[BI] Bojarski, B., and Iwaniec, T., Analytical foundations of the theory of quasiconformal mappings in $\mathbb{R}^n$. *Ann. Acad. Sci. Fenn. Ser. A I Math.* **8** (1983), 257–324.

[Bo] Bonk, M., Uniformization of Sierpiński carpets in the plane. In preparation.

[BHR] Bonk, M., Heinonen, J., and Rohde, S., Doubling conformal densities. *J. Reine Angew. Math.* **541** (2001), 117–141.

[BK1] Bonk, M., and Kleiner, B., Quasisymmetric parametrizations of two-dimensional metric spheres. *Invent. Math.* **150** (2002), 127–183.

[BK2] Bonk, M., and Kleiner, B., Rigidity for quasi-Möbius group actions. *J. Differential Geom.* **61** (2002), 81–106.

[BK3] Bonk, M., and Kleiner, B., Rigidity for quasi-Fuchsian actions on negatively curved spaces. *Internat. Math. Res. Notices* **2004** (61) (2004), 3309–3316.

[BK4] Bonk, M., and Kleiner, B., Conformal dimension and Gromov hyperbolic groups with 2-sphere boundary. *Geom. Topol.* **9** (2005), 219–246.

[BK5] Bonk, M., and Kleiner, B., Quasi-hyperbolic planes in hyperbolic groups. *Proc. Amer. Math. Soc.* **133** (2005), 2491–2494.

[BK6] Bonk, M., and Kleiner, B., Uniformization of metric Sierpiński carpets. In preparation.

[BKM] Bonk, M., Kleiner, B., and Merenkov, S., Rigidity of Schottky sets. Preprint, 2006.

[BMe] Bonk, M., and Merenkov, S., Quasisymmetric rigidity of Sierpiński carpets. In preparation.

[BMy] Bonk, M., and Meyer, D., Topological rational maps and subdivisions. In preparation.

[BS] Bonk, M., and Schramm, O., Embeddings of Gromov hyperbolic spaces. *Geom. Funct. Anal.* **10** (2000), 266–306.

[BP] Bourdon, M., and Pajot, H., Cohomologie l_p et espaces de Besov. *J. Reine Angew. Math.* **558** (2003), 85–108.

[Bow] Bowditch, B. H., A topological characterisation of hyperbolic groups. *J. Amer. Math. Soc.* **11** (1998), 643–667.

[BBI] Burago, D., Burago, Y., and Ivanov, S., *A course in metric geometry*. Grad. Stud. Math. 33, Amer. Math. Soc., Providence, RI, 2001.

[Ca] Cannon, J. W., The combinatorial Riemann mapping theorem. *Acta Math.* **173** (1994), 155–234.

[CFKP] Cannon, J. W., Floyd, W. J., Kenyon, R., and Parry, W. R., Constructing rational maps from subdivision rules. *Conform. Geom. Dyn.* **7** (2003), 76–102.

[CFP] Cannon, J. W., Floyd, W. J., and Parry, W. R., Finite subdivision rules. *Conform. Geom. Dyn.* **5** (2001), 153–196.

[Co] Coornaert, M., Mesures de Patterson-Sullivan sur le bord d'un espace hyperbolique au sens de Gromov. *Pacific J. Math.* **159** (1993), 241–270.

[DS] David, G., and Semmes, S., *Fractured fractals and broken dreams*. Oxford Lecture Ser. Math. Appl. 7, The Clarendon Press, Oxford University Press, New York 1997.

[DH] Douady, A., and Hubbard, J. H., A proof of Thurston's topological characterization of rational functions. *Acta Math.* **171** (1993), 263–297.

[GH] Ghys, E., and de la Harpe, P. (eds.), *Sur les Groupes Hyperboliques d'après Mikhael Gromov*. Progr. Math. 83, Birkhäuser, Boston 1990.

[Gr] Gromov, M., Hyperbolic Groups. In *Essays in Group Theory* (ed. by S. Gersten), Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–265.

[He] Heinonen, J., *Lectures on Analysis on Metric Spaces*. Universitext, Springer-Verlag, New York 2001.

[HK] Heinonen, J., and Koskela, P., Quasiconformal maps in metric spaces with controlled geometry. *Acta Math.* **181** (1998), 1–61.

[HKST] Heinonen, J., Koskela, P., Shanmugalingam, N., and Tyson, J. T., Sobolev classes of Banach space-valued functions and quasiconformal mappings. *J. Anal. Math.* **85** (2001), 87–139.

[HS] Heinonen, J., and Semmes, S., Thirty-three yes or no questions about mappings, measures, and metrics. *Conform. Geom. Dyn.* **1** (1997), 1–12.

[H-R] Hurdal, M. K., Bowers, P. L., Stephenson, K., Sumners, D. W. L., Rehm, K., Schaper, K., and Rottenberg, D. A., Quasi-conformally flat mapping the human cerebellum. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99* (ed. by C. Taylor and A. Colchester), Lecture Notes in Comput. Sci. 1679, Springer-Verlag, Berlin 1999, 279–286.

[KB] Kapovich, I., and Benakli, N., Boundaries of hyperbolic groups. In *Combinatorial and geometric group theory* (New York, 2000/Hoboken, NJ, 2001), Contemp. Math. 296, Amer. Math. Soc., Providence, RI, 2002, 39–93.

[KK] Kapovich, M., and Kleiner, B., Hyperbolic groups with low-dimensional boundary. *Ann. Sci. École Norm. Sup.* (4) **33** (2000), 647–669.

[KL] Keith, S., and Laakso, T. J., Conformal Assouad dimension and modulus. *Geom. Funct. Anal.* **14** (2004), 1278–1321.

[KR1] Korányi, A., and Reimann, H. M., Quasiconformal mappings on the Heisenberg group. *Invent. Math.* **80** (1985), 309–338.

[KR2] Korányi, A., and Reimann, H. M., Foundations for the theory of quasiconformal mappings on the Heisenberg group. *Adv. Math.* **111** (1995), 1–87.

[Kü] Kühnau, R. K., Herbert Grötzsch zum Gedächtnis. *Jahresber. Deutsch. Math.-Verein.* **99** (1997), 122–145.

[My] Meyer, D., Quasisymmetric embedding of self similar surfaces and origami with rational maps. *Ann. Acad. Sci. Fenn. Math.* **27** (2002), 461–484.

[Mo] Mostow, G. D., *Strong rigidity of locally symmetric spaces*. Ann. of Math. Stud. 78, Princeton University Press, Princeton, NJ., 1970; University of Tokyo Press, Tokyo 1973.

[Pa1] Pansu, P., Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rang un. *Ann. of Math.* (2) **129** (1989), 1–60.

[Pa2] Pansu, P., Dimension conforme et sphère à l'infini des variétés à courbure négative. *Ann. Acad. Sci. Fenn. Ser. A I Math.* **14** (1989), 177–212.

[Pau] Paulin, F., Un groupe hyperbolique est déterminé par son bord. *J. London Math. Soc.* (2) **54** (1996), 50–74.

[Re] Reshetnyak, Yu. G., *Space mappings with bounded distortion*. Transl. Math. Monogr. 73, Amer. Math. Soc., Providence, RI, 1989.

[Ri] Rickman, S., *Quasiregular Mappings*. Ergeb. Math. Grenzgeb. 26, Springer-Verlag, Berlin, Heidelberg, New York 1993.

[Sc] Schramm, O., Transboundary extremal length. *J. Anal. Math.* **66** (1995), 307–329.

[Se1] Semmes, S., On the nonexistence of bi-Lipschitz parameterizations and geometric problems about A_∞-weights. *Rev. Mat. Iberoamericana* **12** (1996), 337–410.

[Se2] Semmes, S., Good metric spaces without good parameterizations. *Rev. Mat. Iberoamericana* **12** (1996), 187–275.

[Su1] Sullivan, D., On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions. In *Riemann surfaces and related topics*, Proceedings of the 1978 Stony Brook Conference, Ann. of Math. Stud. 97, Princeton University Press, Princeton, NJ, 1981, 465–496

[Su2] Sullivan, D., Discrete conformal groups and measurable dynamics. *Bull. Amer. Math. Soc.* (*N.S.*) **6** (1982), 57–73.

[Tu] Tukia, P., On quasiconformal groups. *J. Analyse Math.* **46** (1986), 318–346.

[TV] Tukia, P., and Väisälä, J., Quasisymmetric embeddings of metric spaces. *Ann. Acad. Sci. Fenn. Ser. A I Math.* **5** (1980), 97–114.

[Ty] Tyson, J. T., Quasiconformality and quasisymmetry in metric measure spaces. *Ann. Acad. Sci. Fenn. Math.* **23** (1998), 525–548.

[Vä1] J. Väisälä, *Lectures on n-dimensional quasiconformal mappings*. Lecture Notes in Math. 229, Springer-Verlag, Berlin, Heidelberg, New York 1971.

[Vä2] Väisälä, J., Quasi-Möbius maps. *J. Analyse Math.* **44** (1984/85), 218–234.

[Vä3] Väisälä, J., Quasisymmetric maps of products of curves into the plane. *Rev. Roumaine Math. Pures Appl.* **33** (1988), 147–156.

[Wi] Wildrick, K., Quasisymmetric parametrizations of two-dimensional metric planes. Preprint, 2006.

[Wh] Whyburn, G. T., Topological characterization of the Sierpiński curve. *Fund. Math.* **45** (1958), 320–324.

[Yu] Yue, Ch., Dimension and rigidity of quasi-Fuchsian representations. *Ann. of Math.* (2) **143** (1996), 331–355.

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
E-mail: mbonk@umich.edu

Local *Tb* theorems and applications in PDE

Steve Hofmann*

Abstract. A *Tb* theorem is a boundedness criterion for singular integrals, which allows the L^2 boundedness of a singular integral operator T to be deduced from sufficiently good behavior of T on some suitable non-degenerate test function b. However, in some PDE applications, including, for example, the solution of the Kato problem for square roots of divergence form elliptic operators, it may be easier to test the operator T locally (say on any given dyadic cube Q), on a test function b_Q that depends upon Q, rather than on a single, globally defined b. Or to be more precise, in the applications, it may be easier to find a family of b_Q's for which Tb_Q is locally well behaved, than it is to find a single b for which Tb is nice globally. In this lecture, we shall discuss some versions of local *Tb* theorems, as well as some applications to PDE.

Mathematics Subject Classification (2000). Primary 42B20, 42B25; Secondary 35J25, 35J55, 47F05, 47B44.

Keywords. *Tb* theorem, singular integrals, square functions, boundary value problems, Kato problem, layer potentials.

1. Introduction

The *Tb* theorem, and its predecessor, the $T1$ theorem, were introduced in large part to better understand the Cauchy integral operator on a Lipschitz curve, and the related Calderón commutators. In this note, we shall discuss more recent "local" versions of the *Tb* theorem, as well as the application of such theorems to some questions in PDE.

We begin by recalling the statements of the original $T1$ and *Tb* theorems. To this end, we require a few definitions.

We say that T is a singular integral operator (in the generalized sense of Coifman and Meyer), if T is a mapping from test functions $\mathcal{D}(\mathbb{R}^n)$ into distributions $\mathcal{D}'(\mathbb{R}^n)$, which is associated to a "Calderón–Zygmund kernel" $K(x, y)$, in the sense that

$$\langle T\varphi, \psi\rangle = \int\int \psi(x)K(x, y)\varphi(y)dxdy,$$

whenever $\varphi, \psi \in C_0^\infty(\mathbb{R}^n)$ with disjoint supports (the theory can be extended to settings other than Euclidean space, and there are worthwhile reasons for doing so,

*supported by NSF

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

but we shall not insist much on that point, for the sake of simplicity of exposition). A Calderón–Zygmund kernel is one which satisfies the "standard" bounds

$$|K(x, y)| \leq C|x - y|^{-n} \tag{1.1}$$

and

$$|K(x, y + h) - K(x, y)| + |K(x + h, y) - K(x, y)| \leq \frac{C|h|^{\alpha}}{|x - y|^{n+\alpha}}, \tag{1.2}$$

for some $\alpha \in (0, 1]$, whenever $|x - y| \geq 2|h|$. A singular integral operator T is said to satisfy the "Weak Boundedness Property" (WBP), if

$$\sup \left(R^{-n}|\langle T\varphi, \psi\rangle|\right) \leq C < \infty, \tag{1.3}$$

where the supremum runs over all $R > 0$, over all balls $B(x, R)$ of radius R and arbitrary center x, and over all test functions φ, ψ supported in $B(x, R)$, and normalized so that $\|\varphi\|_{\infty} + R\|\nabla\varphi\|_{\infty} \leq 1$ and $\|\psi\|_{\infty} + R\|\nabla\psi\|_{\infty} \leq 1$. In order to demystify this condition, we note that it holds automatically for any L^2 bounded operator (just apply Cauchy–Schwarz). Moreover, with just a small amount of work, it can be shown that, given an anti-symmetric kernel (i.e., one for which $K(x, y) = -K(y, x)$), which in addition satisfies the size condition (1), there is an associated "principal value" type singular integral operator for which WBP holds.

We recall that BMO is the space of locally integrable functions modulo constants for whom the norm

$$\|b\|_{*} = \sup |Q|^{-1} \int_{Q} |b(x) - [b]_{Q}|dx$$

is finite. Here, the supremum runs over all cubes (balls work just as well) with sides parallel to the co-ordinate axes, and

$$[b]_{Q} \equiv |Q|^{-1} \int_{Q} b(x)dx.$$

The $T1$ theorem of David and Journé [DJ] is the following:

Theorem 1.1. *Suppose that T is a singular integral operator associated to a standard kernel $K(x, y)$ satisfying* (1) *and* (2). *Then T extends to a bounded operator on L^2 if and only if T satisfies WBP, and $T1$, $T^*1 \in$ BMO.*

Here, T^* is the formal transpose of T. It is of course associated to the kernel $K^*(x, y) = K(y, x)$. One might ask whether T and T^* are well defined on constant functions, but it is not hard to show, using the kernel condition (2), that $T1$ and T^*1 exist as distributions modulo constants. This result may be understood as follows. If T is bounded on L^2, and its kernel satisfies the smoothness condition (1.2), then by a result obtained independently by Peetre [P], Spanne [Sp] and Stein [St] it follows

that $T: L^\infty \to \mathrm{BMO}$, and similarly for T^*. Conversely, if both T and T^* are bounded from $L^\infty \to \mathrm{BMO}$, then by duality and interpolation (using results of Fefferman and Stein [FS]), we have that T is bounded on L^2. The $T1$ theorem says that in order to obtain the latter conclusion, one need not test T on all of L^∞, but rather, only on a very special element of L^∞, namely the constant function 1.

The Tb theorem is an extension of the $T1$ theorem, in which the function 1 is replaced by a suitable function $b \in L^\infty$ (or, more generally, by two such functions b_1 and b_2: one for T, and one for T^*). One supposes that $b_2 T b_1$ is a mapping from test functions to distributions which satisfies WBP (in particular, principal value operators associated to anti-symmetric kernels have this property) and that $Tb_1, T^*b_2 \in \mathrm{BMO}$. Then, if b_1 and b_2 are sufficiently non-degenerate, one again deduces that T extends to a bounded operator on L^2. In the original versions of this theorem, b was assumed to be essentially bounded and "accretive", i.e., for some $\delta > 0$,

$$\Re e\ b \geq \delta,$$

or merely "pseudo-accretive":

$$\inf_Q \left|[b]_Q\right| \geq \delta,$$

or even "para-accretive", a relaxed version of pseudo-accretivity in which nondegeneracy of the average over each given cube is replaced by nondegeneracy of the average over some sub-cube of comparable size. The special case that $Tb_1 = 0 = T^*b_2$ (here 0 is meant in the sense of BMO, i.e., modulo constants) and b_1, b_2 are accretive, is due to McIntosh and Meyer [McM], the general case to David, Journé and Semmes [DJS].

The special case treated in [McM] already had a spectacular application: as a direct corollary, one obtains an alternative proof of the Cauchy integral theorem of Coifman, McIntosh and Meyer. Indeed, for a Lipschitz function A, the kernel $(x-y+i(A(x)-A(y))^{-1}$ is anti-symmetric and standard, so that L^2 boundedness of

$$T_A f(x) = p.v.\ \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{f(y)}{x - y + i(A(x) - A(y)} \, dy$$

follows immediately from the theorem of [McM] and the observation that, at least formally, by the formula of Plemelj, $T_A b = 1/2$, where $b = 1 + iA'$. In practice, some care must be taken in interpreting the Plemelj formula on an infinite graph, but this can be managed.

In some applications, it may not be at all evident that there is an accretive (or pseudo-accretive) b for which Tb is well behaved. On the other hand, in such cases it is sometimes possible to find a family $\{b_Q\}$, indexed by dyadic cubes Q, such that Tb_Q behaves well locally on Q. This motivates the introduction of the notion of a "local Tb theorem", in which good local control of a singular integral operator T, on each member of a family of suitably non-degenerate functions b_Q (one for each dyadic cube Q), still suffices to deduce L^2 boundedness of T. The appropriate version of

non-degeneracy in this setting was introduced by M. Christ [Ch]: a "pseudo-accretive system" is a collection of functions $\{b_Q\}$, indexed on the dyadic cubes, with b_Q supported in Q and integrable, such that for some $\delta > 0$, we have that

$$\left| |Q|^{-1} \int_Q b_Q \right| \geq \delta. \tag{1.4}$$

The first local *Tb* theorem was proved by Christ:

Theorem 1.2 ([Ch]). *Suppose that T is a singular integral operator associated to a standard kernel K(x,y), which in addition we assume to be in L^∞. Suppose also that there are constants $\delta > 0$ and $C_0 < \infty$, and pseudo-accretive systems $\{b^1_Q\}$, $\{b^2_Q\}$, with* $\operatorname{supp} b^i_Q \subseteq Q$, $i = 1, 2$, *such that for each dyadic cube Q,*

(i) $\|b^1_Q\|_{L^\infty(Q)} + \|b^2_Q\|_{L^\infty(Q)} \leq C_0$,

(ii) $\|Tb^1_Q\|_{L^\infty(Q)} + \|T^* b^2_Q\|_{L^\infty(Q)} \leq C_0$,

(iii) $\delta|Q| \leq \min(\left|\int_Q b^1_Q\right|, \left|\int_Q b^2_Q\right|)$.

Then T extends to a bounded operator on L^2, with bound depending only on n, δ, C_0 and the kernel constants in (1.1) *and* (1.2)*, but not on the L^∞ norm of $K(x, y)$.*

A few remarks are in order. The assumption that $K \in L^\infty$ is merely qualitative, and is satisfied, e.g., by smooth truncations of a standard kernel. This assumption allows one to make certain formal manipulations with impunity, during the course of the proof. Christ actually proved this theorem in the setting of a space of homogeneous type (that is, a space endowed with a pseudo-metric and a doubling measure), which (as he demonstrated) possesses a suitable version of a "dyadic cube" structure. Christ's theorem and the technique of its proof are related to the solution of Painlevé's problem concerning the characterization of those compact sets $K \subset \mathbb{C}$ for which there exist non-constant bounded analytic functions on $\mathbb{C} \setminus K$. We will not discuss this aspect of the theory in detail, but we mention that extensions of either local or global *Tb* theorems to the non-doubling setting have been obtained by G. David [D1] and by Nazarov, Treil and Volberg [NTV1], [NTV2]; moreover, the circle of ideas involved in [Ch], [D1] and [NTV1], [NTV2] have played a crucial role in the solution of the Painlevé problem, see Mattila, Melnikov and Verdera [MMV], G. David [D1], [D2] and X. Tolsa [T], and also Volberg [Vo], where the higher dimensional version of this theory is treated. See also the article of Tolsa in these proceedings.

Instead, in this note we shall concentrate on extensions of Christ's result in another direction, in which L^∞ control of b_Q and Tb_Q is replaced by local, scale invariant L^2 control. These extensions have been useful in certain applications in PDE, including the solution of the Kato problem. In the next two sections, we discuss local *Tb* theorems for square functions, and for singular integrals, respectively.

Acknowledgements. Various portions of the material discussed in this paper were joint work with M. Alfonseca, P. Auscher, A. Axelsson, M. Lacey, S. Kim, A. McIntosh, C. Muscalu, T. Tao, P. Tchamitchian and C. Thiele. To all of them I express my deepest appreciation.

2. Local *Tb* theorems for square functions and applications

We begin with a local *Tb* theorem for square functions, which extends a global version due to Semmes [S]. Suppose that we have a family of kernels $\{\psi_t(x, y)\}_{t\in(0,\infty)}$, satisfying, for some exponent $\alpha > 0$,

$$\begin{aligned} |\psi_t(x, y)| &\le C\frac{t^\alpha}{(t + |x - y|)^{n+\alpha}}, \\ |\psi_t(x, y + h) - \psi_t(x, y)| &\le C\frac{|h|^\alpha}{(t + |x - y|)^{n+\alpha}} \end{aligned} \tag{2.1}$$

whenever $|h| \le \frac{1}{2}|x - y|$ or $|h| \le |t|/2$.

Theorem 2.1. *Let $\theta_t f(x) \equiv \int \psi_t(x, y) f(y)\, dy$, where $\psi_t(x, y)$ satisfies* (2.1)*. Suppose also that there exist constants $\delta > 0$, $C_0 < \infty$, and a system $\{b_Q\}$ of functions indexed by dyadic cubes $Q \subseteq \mathbb{R}^n$ such that for each dyadic cube Q*

(i) $\int_{\mathbb{R}^n} |b_Q|^2 \le C_0|Q|$,

(ii) $\int_0^{\ell(Q)}\int_Q |\theta_t b_Q(x)|^2 \frac{dxdt}{t} \le C_0|Q|$,

(iii) $\delta|Q| \le \left|\int_Q b_Q\right|$.

Then we have the square function bound

$$\iint_{\mathbb{R}^{n+1}_+} |\theta_t f(x)|^2 \frac{dxdt}{t} \le C\|f\|_2^2.$$

Proof. The proof combines the ideas of [S] and [AT] with those of [Ch], [HMc], [HLMc] and [AHLMcT]. (Actually, the argument below preceded the subsequent matrix-valued versions used in [HMc], [HLMc] and [AHLMcT] to solve the Kato problem, but the author never published it in this scalar-valued form; see also Auscher's lecture notes on the Kato problem [A], where the present formulation is given explicitly). We begin by recalling the following well-known fact, due explicitly to Christ and Journé [CJ], but also at least implicit in the work of Coifman and Meyer [CM].

Proposition 2.2 ([CJ])**.** *Let $\theta_t f(x) \equiv \int_{\mathbb{R}^n} \psi_t(x, y) f(y)dy$, where $\psi_t(x, y)$ satisfies* (2.1)*. Suppose that we have the Carleson measure estimate*

$$\sup_Q \frac{1}{|Q|}\int_0^{\ell(Q)}\int_Q |\theta_t 1(x)|^2 \frac{dxdt}{t} \le C. \tag{2.2}$$

Then we have the square function estimate

$$\iint_{\mathbb{R}^{n+1}_+} |\theta_t f(x)|^2 \, \frac{dxdt}{t} \leq C\|f\|_2^2. \tag{2.3}$$

Remark 2.3. The converse direction (i.e. that (2.3) implies (2.2)) is essentially due to Fefferman and Stein [FS].

Thus, to prove Theorem 2.1, it suffices to verify that $|\theta_t 1|^2 \, dxdt/t$ is a Carleson measure, given the existence of a family $\{b_Q\}$ satisfying hypotheses (i), (ii) and (iii) of the theorem. To this end, we first observe that, as in [S] and [AT], it is enough to verify that for $\{b_Q\}$ as in the theorem, we have the bound

$$\sup_Q \frac{1}{|Q|} \iint_{R_Q} |\theta_t 1|^2 \, \frac{dxdt}{t} \leq C \sup_Q \frac{1}{|Q|} \iint_{R_Q} |(\theta_t 1)(P_t b_Q)|^2 \, \frac{dxdt}{t} + C, \tag{2.4}$$

where $R_Q \equiv Q \times (0, \ell(Q))$ is the "Carleson box" above Q, and where P_t is a nice approximate identity, whose kernel satisfies, say, (2.1). Indeed, suppose momentarily that (2.4) holds. Then to obtain (2.2) (and thus also the conclusion of the theorem), it suffices to show that the right hand side of (2.4) is bounded. Following [CM], we write

$$\begin{aligned}(\theta_t 1) P_t b_Q &= [(\theta_t 1) P_t b_Q - \theta_t b_Q] + \theta_t b_Q \\ &= R_t b_Q + \theta_t b_Q.\end{aligned}$$

The contribution of $\theta_t b_Q$ is bounded, by hypothesis (ii) of the theorem. Moreover, by (2.1) and the fact that $R_t 1 = 0$, it follows by a well-known orthogonality argument that

$$\iint_{\mathbb{R}^{n+1}_+} |R_t f(x)|^2 \, \frac{dxdt}{t} \leq C\|f\|_2^2.$$

Thus, by (i), the contribution of $R_t b_Q$ is also bounded.

Therefore, to finish the proof of the theorem, it remains to verify (2.4). In fact, it suffices to proof that (2.4) holds with P_t replaced by the dyadic averaging operator A_t, defined by

$$A_t f(x) \equiv A_t^Q f(x) \equiv \frac{1}{|Q(x,t)|} \int_{Q(x,t)} f(y)\, dy,$$

where $Q(x,t)$ denotes the minimal dyadic subcube of Q containing x, with side length at least t. Indeed, a standard orthogonality argument yields the fact that

$$\iint_{R_Q} |(A_t - P_t) f(x)|^2 \, \frac{dxdt}{t} \leq C\|f\|_2^2,$$

so that the error is bounded.

Now, by a well-known "John–Nirenberg" type lemma for Carleson measures (see, e.g. [AHLT, Lemma 3.3]), in order to establish (2.6) (or rather its analogue with A_t

in place of P_t), it suffices to show that there is a positive constant $\eta > 0$ such that for each Q, there is a dyadic sawtooth region

$$E_Q^* \equiv R_Q \setminus (\cup R_{Q_j}), \tag{2.5}$$

where $\{Q_j\}$ are non-overlapping dyadic sub-cubes of Q, with

$$|Q \setminus (\cup Q_j)| > \eta |Q|$$

and

$$\iint_{E_Q^*} |\theta_t 1(x)|^2 \, \frac{dxdt}{t} \leq C \iint_{E_Q^*} |(\theta_t 1(x))(A_t b_Q(x))|^2 \, \frac{dxdt}{t}. \tag{2.6}$$

We establish (2.6) via a stopping time argument as in [HMc], [HLMc] and [AHLMcT] (but see also [Ch], where a similar idea had previously appeared). Our starting point is (iii). Dividing by an appropriate complex constant, we may suppose that

$$\frac{1}{|Q|} \int_Q b_Q = 1. \tag{2.7}$$

We then sub-divide Q dyadically, to select a family of non-overlapping cubes $\{Q_j\}$ which are maximal with respect to the property that

$$\Re e \, \frac{1}{|Q_j|} \int_{Q_j} b_Q \leq 1/2. \tag{2.8}$$

If E_Q^* is defined as in (2.5) with respect to this family $\{Q_j\}$, then by construction, if $(x, t) \in E_Q^*$, it follows that

$$\frac{1}{2} \leq \Re e \, A_t b_Q(x),$$

so that (2.6) holds with $C = 4$. It remains only to verify that there exists $\eta > 0$ such that

$$|E| > \eta |Q|, \tag{2.9}$$

where $E \equiv Q \setminus (\cup Q_j)$. By (2.7) we have that

$$\begin{aligned} |Q| = \int_Q b_Q = \Re e \int_Q b_Q = \Re e \int_E b_Q + \Re e \sum_j \int_{Q_j} b_Q \\ \leq |E|^{\frac{1}{2}} \left(\int_Q |b_Q|^2 \right)^{\frac{1}{2}} + \frac{1}{2} \sum |Q_j|, \end{aligned}$$

when in the last step we have used (2.8). From hypothesis (i) of Theorem 2.1, we then obtain that

$$|Q| \leq C |E|^{\frac{1}{2}} |Q|^{\frac{1}{2}} + \frac{1}{2} |Q|,$$

and (2.9) now follows readily. This concludes the proof of Theorem 2.1. □

As alluded to above, the previous theorem has an extension to the matrix valued setting. We shall explain momentarily why this is interesting. Let $\mathbb{M}^N$ denote the space of $N \times N$ matrices with complex entries.

Theorem 2.4. *Suppose that* $\mathbf{\Psi}_t : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}^N$ *satisfies* (2.1). *For* $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{C}^N$ *define the operator*

$$\mathbf{\Theta}_t \cdot \boldsymbol{f}(x) \equiv \int \mathbf{\Psi}_t(x, y) \cdot \boldsymbol{f}(y)\, dy. \tag{2.10}$$

Suppose also that there are constants $\delta > 0$, $C_0 < \infty$ *and a system of matrix valued functions* $\boldsymbol{b}_Q : \mathbb{R}^n \to \mathbb{M}^N$, *indexed on the dyadic cubes, such that*

(i) $\int_{\mathbb{R}^n} |\boldsymbol{b}_Q|^2 \leq C_0 |Q|$,

(ii) $\int_0^{\ell(Q)} \int_Q |\mathbf{\Theta}_t \boldsymbol{b}_Q(x)|^2 \, \frac{dxdt}{t} \leq C_0 |Q|$,

(iii) $\delta |\xi|^2 \leq \Re e\ \xi \cdot \big(|Q|^{-1} \int_Q \boldsymbol{b}_Q\big)\overline{\xi}$.

where the ellipticity condition (iii) *holds for all* $\xi \in \mathbb{C}^N$, *and where the action of* $\mathbf{\Theta}_t$ *on the matrix valued function* $\boldsymbol{b}_Q$ *is defined in the obvious way as in* (2.10) *by viewing the kernel* $\mathbf{\Psi}_t(x, y)$ *as a* $1 \times N$ *matrix which multiplies the* $N \times N$ *matrix* $\boldsymbol{b}_Q$. *Then*

$$\iint_{\mathbb{R}^{n+1}_+} |\mathbf{\Theta}_t \cdot \boldsymbol{f}|^2 \, \frac{dxdt}{t} \leq C \|\boldsymbol{f}\|_2^2.$$

It turns out that a variant of this theorem lies at the heart of the solution of the Kato problem [HMc], [HLMc], [AHLMcT] (see also [AT]). We now sketch the proof, which is essentially the same as the argument used to establish the Kato conjecture. Let $\mathbf{1}$ denote the $N \times N$ identity matrix. Since

$$\mathbf{\Theta}_t \mathbf{1} = (\theta_t^1 1, \theta_t^2 1, \ldots, \theta_t^N 1),$$

Proposition 2.3 therefore implies that it is enough to show that $|\mathbf{\Theta}_t \mathbf{1}|^2 \, t^{-1} dxdt$ is a Carleson measure. For ε small, but fixed, we cover $\mathbb{C}^N$ by cones of aperture ε. Enumerating these cones as $\Gamma_1^\varepsilon, \ldots, \Gamma_K^\varepsilon,$, where $K = K(\varepsilon, N)$, we see that

$$\int_0^{\ell(Q)} \int_Q |\mathbf{\Theta}_t \mathbf{1}|^2 \, \frac{dxdt}{t} = \sum_{k=1}^K \int_0^{\ell(Q)} \int_Q |\mathbf{\Theta}_t \mathbf{1}|^2 \, 1_{\Gamma_k^\varepsilon}(\mathbf{\Theta}_t \mathbf{1}) \, \frac{dxdt}{t}.$$

Thus, it is enough to show that there is a uniform constant

$$C_1 = C_1(\varepsilon, \delta, C_0, n, N)$$

such that, for each fixed cone Γ^ε with ε small enough,

$$\sup_Q |Q|^{-1} \int_0^{\ell(Q)} \int_Q |\mathbf{\Theta}_t \mathbf{1}|^2 \, 1_{\Gamma^\varepsilon}(\mathbf{\Theta}_t \mathbf{1}) \, \frac{dxdt}{t} \leq C_1.$$

To this end, normalizing so that $\delta = 1$, and fixing Q, we follow the stopping time argument of the previous theorem, in the present case extracting dyadic subcubes $Q_j \subset Q$ which are maximal with respect to the property that at least one of the following holds:

$$\int_{Q_j} |\boldsymbol{b}_Q| \geq \frac{1}{4\varepsilon} \tag{2.11}$$

or

$$\Re e\, \nu \cdot \left(|Q_j|^{-1} \int_{Q_j} \boldsymbol{b}_Q \right) \overline{\nu} \leq \frac{3}{4}, \tag{2.12}$$

where $\nu \in \mathbb{C}^N$ is the unit normal in the direction of the central axis of Γ^ε, i.e.,

$$\Gamma^\varepsilon = \left\{ z \in \mathbb{C}^N : \left| \frac{z}{|z|} \right| - \nu \right| < \varepsilon \right\}.$$

As in the proof of the previous theorem, one may check that

$$|E| \equiv |Q \setminus (\cup Q_j)| \geq \eta |Q|,$$

for some fixed $\eta > 0$. Moreover, for $(x,t) \in E_Q^* \equiv R_Q \setminus \left(\bigcup R_{Q_j} \right)$, (we recall that $R_Q \equiv Q \times (0, \ell(Q)$ is the Carleson box above Q) and for $z \in \Gamma^\varepsilon$, we claim that

$$|z \cdot A_t \boldsymbol{b}_Q(x) \overline{\nu}| \geq \frac{1}{2} |z|, \tag{2.13}$$

where again A_t denotes the dyadic averaging operator with respect to the dyadic grid of Q. Indeed, since the opposite inequalities to (2.14) and (2.15) hold in E_Q^*, we have that

$$|\omega \cdot A_t \boldsymbol{b}_Q(x) \overline{\nu}| \geq |\nu \cdot A_t \boldsymbol{b}_Q(x) \overline{\nu}| - |(\omega - \nu) \cdot A_t \boldsymbol{b}_Q(x) \overline{\nu}| \geq \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$$

whenever $|\omega - \nu| < \varepsilon$ and $(x,t) \in E_Q^*$. Taking $\omega = z/|z|$, with $z \in \Gamma^\varepsilon$, we obtain (2.16).

Consequently, we have that

$$\iint_{E_Q^*} |\boldsymbol{\Theta}_t \mathbf{1}|^2 \, 1_{\Gamma^\varepsilon}(\boldsymbol{\Theta}_t \mathbf{1}) \, \frac{dxdt}{t} \leq 4 \iint_{E_Q^*} |\boldsymbol{\Theta}_t \mathbf{1} \cdot A_t \boldsymbol{b}_Q \overline{\nu}|^2 \, \frac{dxdt}{t},$$

and the rest of the proof follows as in the previous theorem.

As mentioned above, a variant of this last theorem leads to the solution of the Kato problem. We recall the statement of the problem. Let B be an $n \times n$ matrix of complex, L^∞ coefficients, defined on $\mathbb{R}^n$, and satisfying the ellipticity (or "accretivity") condition

$$\lambda |\xi|^2 \leq \Re e \, \langle B\xi, \xi \rangle \equiv \Re e \sum_{i,j} B_{ij}(x) \, \xi_j \, \overline{\xi}_i, \quad \|B\|_\infty \leq \Lambda, \tag{2.14}$$

for $\xi \in \mathbb{C}^n$ and for some λ, Λ such that $0 < \lambda \leq \Lambda < \infty$. We define a divergence form operator

$$Ju \equiv -\operatorname{div}(B(x)\nabla u), \tag{2.15}$$

which we interpret in the usual weak sense via a sesquilinear form. The accretivity condition (2.14) enables one to define an accretive square root $\sqrt{J} \equiv J^{1/2}$ (see [K1], [K2]), and the "Kato problem", or "square root problem", is to establish the estimate

$$\|\sqrt{J} f\|_{L^2(\mathbb{R}^n)} \leq C\|\nabla f\|_{L^2(\mathbb{R}^n)}, \tag{2.16}$$

with C depending only on n, λ and Λ. The latter estimate is connected with the question of the analyticity of the mapping $B \to J^{\frac{1}{2}}$, which in turn has applications to the perturbation theory for certain classes of hyperbolic equations (see [Mc]). We remark that (2.16) is equivalent to the opposite inequality for the square root of the adjoint operator J^* (which amounts to the L^2 boundedness of the Riesz transforms $\nabla(J^*)^{-1/2}$). In [HMc], [HLMc], [AHLMcT], (but see also [AT]), estimate (2.16) was deduced, in effect, from a variant of Theorem 2.12, with $N = n$, in which the matrix $\boldsymbol{b}_Q$ is the derivative matrix of a carefully chosen $\mathbb{C}^n$-valued solution F_Q of an appropriate PDE. For example, one can take F_Q to be a certain $W^{1,2}$ solution of the parabolic equation $\frac{\partial u}{\partial t} + Ju = 0$, with t frozen at the scale $t = (\varepsilon \ell(Q))^2$ (ε chosen small, but fixed depending on n, λ and Λ), or it could be a solution of the resolvent equation $(1 + (\varepsilon \ell(Q))^2 J)F_Q = x$. In the case of the Kato problem, the operators $\boldsymbol{\Theta}_t$ which arise do not satisfy the kernel conditions (2.1), but they do posses some extra structure inherited from the operator J, which suffices to carry through the same argument sketched above in the proof of Theorem 2.12.

3. Local *Tb* theorems for singular integrals and applications

To help motivate our next application, we discuss the Kato problem from the perspective of elliptic boundary value problems. Consider the Dirichlet problem

$$\begin{cases} u_{tt} + \operatorname{div}_x B(x)\nabla_x u = 0 \text{ in } \mathbb{R}^{n+1}_+ = \{(x,t) \in \mathbb{R}^n \times (0,\infty)\} \\ u(x,0) = f(x) \in L^2(\mathbb{R}^n). \end{cases} \tag{D}$$

Then a solution u is given in terms of the Poisson semigroup: $u(x,t) = e^{-t\sqrt{J}} f(x)$. Note that the outward normal derivative is given by

$$\frac{\partial u}{\partial \nu} = -\frac{\partial u}{\partial t} = \sqrt{J} u,$$

and that the tangential gradient is simply

$$\nabla_{\tan} u = \nabla_x u.$$

Thus the Kato estimate (2.16), together with the reverse inequality for the Riesz transforms $\nabla J^{-1/2}$, can be thought of as a "Rellich identity"

$$\|\frac{\partial u}{\partial \nu}\|_2 \approx \|\nabla_{\tan} u\|_2 \tag{3.1}$$

for solutions of the boundary value problem (D). The Rellich identity, in turn, plays a vital role in the solution of the Neumann and regularity problems with L^2 estimates (see, e.g., Jerison and Kenig [JK2], Verchota [V] and Kenig and Pipher [KP]); moreover, a local scale invariant Rellich identity can be used to establish L^2 estimates for solutions of the Dirichlet problem [JK1], [JK3]. We observe that the equation

$$u_{tt} + \operatorname{div}_x B(x)\nabla_x u = 0$$

can be written in the form

$$\operatorname{div}_{x,t} A(x)\nabla_{x,t} u = 0, \tag{3.2}$$

where A is the $(n+1)\times(n+1)$ matrix

$$\left[\begin{array}{c|c} & 0 \\ B & \vdots \\ & 0 \\ \hline 0\cdots 0 & 1 \end{array}\right]. \tag{3.3}$$

The question then naturally arises whether the special "semi-group structure" (3.3) is needed to establish the "Rellich identity" (3.1) (and more generally, to obtain L^2 estimates for the Dirichlet, Neumann and regularity problems). Perhaps one might be able to consider equations of the type (3.2) with a "full" $(n+1)\times(n+1)$ elliptic coefficient matrix $A(x)$ (still independent of the t variable). Indeed, for real symmetric coefficients, this is the case [JK3], [KP], [Ke]. Unfortunately, the analogous statement fails for coefficients which are real, but non-symmetric (let alone complex): solvability with L^2 estimates does not hold in general if the non-symmetry is sufficiently severe [KKPT]. On the other hand, it turns out that the magnitude of the non-symmetry matters. Suppose that $A_1(x)$ is a complex, L^∞ elliptic matrix (in the sense of (2.17), but now with $\xi \in \mathbb{C}^{n+1}$), and that $\|A_1 - A_0\|_{L^\infty(\mathbb{R}^n)} \leq \varepsilon$, where $A_0(x)$ is real, symmetric, L^∞ and elliptic, and where ε depends only on dimension and the ellipticity parameters for A_0. Then the Rellich identity (3.1) holds for solutions of the equation

$$L_1 u = -\operatorname{div} A_1(x)\nabla u = 0$$

(in (3.1), $\frac{\partial u}{\partial \nu}$ now denotes the "co-normal" derivative), and one has solvability with L^2 estimates for the Dirichlet, Neumann and Regularity problems [AAAHK]. The proof entails establishing an analytic perturbation result for the layer potentials associated to operators close to $L_0 = -\operatorname{div} A_0(x)\nabla$, and therefore the first step requires that we obtain L^2 boundedness (and invertibility) of the layer potentials associated to L_0. We remark here that since L_0 has real, symmetric, t-independent coefficients, the

solvability, with L^2 estimates, of the Dirichlet [JK1], [JK3], [Ke, pp. 63–64] and Neumann and Regularity [KP] problems for the equation $L_0 u = 0$ was already known, but the layer potential theory is new, and is used to jump start the perturbation scheme. This first step brings us back to the subject of local *Tb* theorems, this time for singular integrals rather than for square functions. (Actually, the subsequent analytic perturbation step also uses local *Tb* technology, in the spirit of the proof of the Kato problem, but this aspect of the theory is rather involved, and we shall not discuss it here).

The following theorem was (essentially) proved in [AHMTT].

Theorem 3.1. *Let T be a singular integral operator associated to a kernel K satisfying the Calderón–Zygmund kernel conditions* (1.1) *and* (1.2)*, as well as the generalized truncation condition $K(x, y) \in L^\infty(\mathbb{R}^n \times \mathbb{R}^n)$. Suppose also that there exist pseudo-accretive systems $\{b_Q\}$, $\{b_Q^*\}$ such that* $\operatorname{supp} b_Q$, $\operatorname{supp} b_Q^* \subseteq Q$, *and*

(i) $\int_Q (|b_Q|^{2+\varepsilon} + |b_Q^*|^{2+\varepsilon}) \leq C|Q|$, *for some $\varepsilon > 0$,*

(ii) $\int_Q (|Tb_Q|^2 + |T^* b_Q^*|^2) \leq C|Q|$,

(iii) $\frac{1}{C}|Q| \leq \min(|\int_Q b_Q|, |\int_Q b_Q^*|)$.

Then $T: L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$, with bound independent of $\|K\|_\infty$.

Remark. This theorem was proved in [AHMTT] for so-called "perfect dyadic" singular integral operators, which are associated to a kernel for which the smoothness condition (1.2) is replaced by the condition that

$$K(x, y) - K(x', y) = 0, \tag{3.4}$$

whenever $x, x' \in Q$ and $y \in Q^c$, where Q is a dyadic cube. In that case, one can take ε to be 0 in condition (i). The proof of the present theorem is essentially the same as that in [AHMTT] except that one is forced to treat several error terms caused by the decaying tail in (1.2), which are absent when one has the perfect localization (3.5) (see [AAAHK] for details). We will not give the proof of this theorem here (it is a bit complicated), although we shall briefly discuss some of the ideas involved in its proof. First however, we describe how one may use this theorem to deduce boundedness of the layer potentials associated to a divergence form elliptic operator in $\mathbb{R}^{n+1}_+$, with real symmetric t-independent coefficients.

Suppose now that

$$Lu = -\operatorname{div} A(x)\nabla u,$$

is defined in $\mathbb{R}^{n+1} = \{(x, t) \in \mathbb{R}^n \times (-\infty, \infty)\}$ (so that div, ∇ are taken in all $n+1$ variables x and t), where $A(x)$ is real, symmetric, elliptic and L^∞. There is a global fundamental solution

$$\Gamma(x, t, y, s) = \Gamma(x, t - s, y, 0)$$

associated to L, which by De Giorgi–Nash–Moser estimates satisfies

$$|\Gamma(x,t,y,0)| \leq C(|t|+|x-y|)^{1-n} \tag{3.5}$$

$$\left|\frac{\partial}{\partial t}\Gamma(x,t,y,0)\right| \leq C(|t|+|x-y|)^{-n}, \tag{3.6}$$

$$\left|\frac{\partial}{\partial t}\left(\Gamma(x+h,t,y,0)-\Gamma(x,t,y,0)\right)\right| + \left|\frac{\partial}{\partial t}\left(\Gamma(x,t,y+h,0)-\Gamma(x,t,y,0)\right)\right| \leq C\frac{|h|^{\alpha}}{(|t|+|x-y|)^{n+\alpha}}, \tag{3.7}$$

whenever $|h| \leq \frac{1}{2}|x-y|$ or $|h| \leq |t|/2$, for some $\alpha > 0$. We define as usual the single layer potential operator

$$S_t f(x) \equiv \int_{\mathbb{R}^n} \Gamma(x,t,y,0) f(y)\, dy,$$

and also singular integrals

$$T_t f(x) \equiv \frac{\partial}{\partial t} S_t f(x) = \int_{\mathbb{R}^n} \frac{\partial}{\partial t}\Gamma(x,t,y,0) f(y)\, dy.$$

We observe that, by virtue of (3.7) and (3.8), the kernel of the latter operator satisfies the hypotheses of Theorem 3.4, uniformly in $t > 0$. Thus, the estimate

$$\sup_{t>0} \|T_t f\|_{L^2(\mathbb{R}^n)} \leq C\|f\|_2 \tag{3.8}$$

will follow immediately if we can produce pseudo-accretive systems $\{b_Q\}$, $\{b_Q^*\}$ satisfying the conditions (i),(ii) and (iii). Given (3.9), bounds for the tangential derivatives of $S_t f$ will follow from the estimate

$$\|\nabla_x S_t f\|_2 \leq C\|\frac{\partial}{\partial t} S_t f\|_2,$$

which in turn is easily obtained for real, symmetric coefficients via integration by parts (indeed, it is a consequence of the Rellich identity). The invertibility of the appropriate boundary integral operators is also then obtainable from the Rellich identity. Let us now indicate how one may deduce (3.9) from Theorem 3.4. We shall only produce a pseudo-accretive system for the operator T_t; one may treat T_t^* by a transparent variation of the same method. We recall the following fundamental result of Jerison and Kenig [JK3] (see also [Ke, pp. 63–64]):

Theorem 3.2 ([JK3]). *Suppose that $L = -\operatorname{div} A\nabla$, where A is real, symmetric, $(n+1)\times(n+1)$, t-independent, L^∞ and uniformly elliptic. Then the elliptic-harmonic measure associated to L, in the lower half-space $\mathbb{R}^{n+1}_-$, is absolutely continuous with respect to n-dimensional Lebesgue measure on the boundary $\{t = 0\}$. Moreover,*

if $k^{A_Q^-}(y)$ denotes the Poisson kernel, at the point $A_Q^- = (x_Q, -\ell(Q))$, where Q is a cube on the boundary with center x_Q, then we have the scale invariant estimate

$$\int_{\mathbb{R}^n} (k^{A_Q^-}(y))^{2+\varepsilon}\, dy \leq C|Q|^{-1-\varepsilon}, \tag{3.9}$$

for some $\varepsilon > 0$ depending only on dimension and ellipticity.

We now set

$$b_Q \equiv |Q| 1_Q\, k^{A_Q^-}. \tag{3.10}$$

Observe that condition (i) of Theorem 3.4 follows immediately from Theorem 3.10. Moreover (iii) is an immediate consequence of the following well-known estimate of Caffarelli, Fabes, Mortola and Salsa [CFMS] (see also [Ke, Lemma 1.3.2, p. 9]):

$$\int_Q k^{A_Q^-}(y)\, dy \geq \frac{1}{C}. \tag{3.11}$$

It remains to establish condition (ii). We consider first

$$\begin{aligned} \partial_t S_t \tilde{b}_Q(x) &= |Q| \int_{\mathbb{R}^n} \partial_t \Gamma(x, t, y, 0) k^{A_Q^-}(y)\, dy \\ &= |Q| \partial_t \Gamma(x, t, A_Q^-), \end{aligned}$$

where $\tilde{b}_Q$ is defined as in (3.10) (except that we have dropped the indicator function of the cube Q), and where we have used that for $(x, t) \in \mathbb{R}^{n+1}_+$ fixed, the function $\partial_t \Gamma(x, t, \cdot, \cdot)$ solves $Lu = 0$ in $\mathbb{R}^{n+1}_-$. Since $t > 0$, we then have by (3.5) and translation invariance in t that

$$|\partial_t S_t b_Q(x)| \leq C,$$

uniformly in $(x, t) \in \mathbb{R}^{n+1}_+$. It is not hard to use integration by parts, coupled again with the fact that $\partial_t \Gamma(x, t, \cdot, \cdot)$ is a solution in the lower half-space to obtain a similar estimate for $\partial_t S_t(\eta_Q \tilde{b}_Q)(x)$, where $\eta_Q \in C_0^\infty$, $\eta_Q \equiv 1$ on $5Q$, $\operatorname{supp} \eta_Q \leq 6Q$, with $\|\nabla \eta_Q\|_\infty \leq C/\ell(Q)$. One then obtains the L^2 bound (ii) for $\partial_t S_t b_Q$ by using (i) and the kernel estimate (3.7) to handle the error $\partial_t S_t \left((\eta_Q - 1_Q) b_Q\right)(x)$. We omit the details (the interested reader may consult [AAAHK]).

Let us conclude the article by sketching some of the ideas involved in the proof of Theorem 3.4. We begin by trying to mimic, as far as possible, the proof of Theorem 2.2. By the $T1$ theorem, it is enough to show that $T1 \in \mathrm{BMO}$ (we ignore the matter of establishing WBP – it turns out that there is a local version of the $T1$ condition, in which 1 is replaced by 1_Q, that yields weak boundedness also, and in practice, it is this local condition that one establishes). By [FS], it would be enough to verify the Carleson measure estimate

$$\sup_Q |Q|^{-1} \int_0^{\ell(Q)} \int_Q |\Delta_t T1(x)|^2\, \frac{dxdt}{t} < \infty, \tag{3.12}$$

where

$$\Delta_t f(x) \equiv \int t^{-n} \psi\left(\frac{x-y}{t}\right) f(y)\, dy,$$

and $\psi \in C_0^\infty(\{|x| < 1\})$ is radial with

$$\int_0^\infty |\hat{\psi}(t)|^2 \frac{dt}{t} = 1.$$

We attempt to proceed as in the proof of Theorem 2.2, but now with $\Delta_t T$ in place of θ_t. As before, we use conditions (i) and (iii) to produce a dyadic sawtooth region E_Q^*, with $|\partial E_Q^* \cap Q| > \eta |Q|$, on which

$$|\Delta_t T 1| \leq C |(\Delta_t T 1)(P_t b_Q)|$$

(modulo acceptable errors). It is then enough to control

$$\iint_{E_Q^*} |(\Delta_t T 1)(P_t b_Q)|^2 \frac{dx\,dt}{t}.$$

Again following the idea of [CM], we write

$$(\Delta_t T 1)(P_t b_Q) = \left[(\Delta_t T 1)(P_t b_Q) - \Delta_t T b_Q\right] + \Delta_t T b_Q \equiv \Lambda_t b_Q + \Delta_t T b_Q.$$

The contribution of the second summand can be handled using condition (ii) and the boundedness of the square function

$$f \to \left(\int_0^\infty |\Delta_t f|^2 \frac{dt}{t}\right)^{1/2}.$$

It is the first summand which causes problems. In contrast to the situation in Theorem 2.2, in which the kernel of the operator $R_t = (\theta_t 1) P_t - \theta_t$ gave rise to a bounded square function, the contribution of $\Lambda_t b_Q$ is now problematic, owing to the failure of the estimates (2.1) for the kernel of Λ_t. Let us try to isolate the difficulty, by writing

$$\Lambda_t = [(\Delta_t T 1) P_t - \Delta_t T P_t] + \Delta_t T (P_t - I) \equiv \tilde{R}_t + \tilde{\Lambda}_t.$$

Then we can at least handle $\tilde{R}_t$ exactly as we did R_t in Theorem 2.2: it is not hard to show that its kernel satisfies (2.1) (I am cheating a bit here – the bound for the kernel of $\Delta_t T P_t$ uses WBP) and clearly $\tilde{R}_t 1 = 0$. It is the term $\tilde{\Lambda}_t$ which now causes problems. If $T^* 1 = 0$, or even if $T^* 1 \in \mathrm{BMO}$, then one can prove square function bounds for the contribution of $\tilde{\Lambda}_t$ (this is easiest to do when the Littlewood–Paley operators Δ_t have been discretized, and when T is of "perfect dyadic" type; see the "one-sided *Tb* theorem" in [AHMTT]). In the absence of this "one-sided" condition, the idea is to build discretized Δ_t operators which are adapted to b_Q^* (as in [CJS]), to take advantage of the fact that we can control (locally) $T^* b_Q^*$ in place of $T^* 1$. The difficulty is that these adapted Δ_t operators are now well behaved only

in sawtooth regions on which b_Q^* is non-degenerate, and therefore there is a stopping time construction needed just to reach the analogue of (3.14), which now, for a given cube Q, becomes an estimate over a sawtooth adapted to b_Q^*. Morally speaking, one then proceeds more or less as I have described above. In practice, this is a bit delicate. We refer the interested reader to [AHMTT] and [AAAHK] for details.

References

[AAAHK] Alfonseca, M., Auscher, P., Axelsson, A., Hofmann, S., and Kim, S., Stability of layer potentials and L^2 solvability of boundary value problems for divergence form elliptic equations with complex L^∞ coefficients. Preprint.

[A] Auscher, P., *Lectures on the Kato square root problem*. Unpublished lecture notes, 2001.

[AHLMcT] Auscher, P., Hofmann, S., Lacey, M., McIntosh, A., and Tchamitchian, P., The Solution of the Kato Square Root Problem for Second Order Elliptic Operators on $\mathbb{R}^n$. *Ann. of Math.* **156** (2002), 633–654.

[AHLT] Auscher, P., Hofmann, S., Lewis, J. L., Tchamitchian, P., Extrapolation of Carleson measures and the analyticity of Kato's square root operators. *Acta Math.* **187** (2) (2001), 161–190.

[AHMTT] Auscher, P., Hofmann, S., Muscalu, C., Tao, T., Thiele, C., Carleson measures, trees, extrapolation, and $T(b)$ theorems. *Publ. Mat.* **46** (2) (2002), 257–325.

[AT] Auscher, P., and Tchamitchian, Ph., Square root problem for divergence operators and related topics. *Astérisque* **249** (1998), Société Mathématique de France.

[CFMS] Caffarelli, L., Fabes, E., Mortola, S., and Salsa, S., Boundary behavior of nonnegative solutions of elliptic operators in divergence form. *Indiana Univ. Math. J.* **30** (4) (1981), 621–640.

[Ch] Christ, M., A T(b) theorem with remarks on analytic capacity and the Cauchy integral. *Colloq. Math.* **60/61** (1990), 601–628.

[CJ] Christ, M., and Journé, J.-L., Polynomial growth estimates for multilinear singular integral operators. *Acta Math.* **159** (1–2) (1987), 51–80.

[CJS] Coifman, R., Jones, P., and Semmes, S., Two elementary proofs of the L^2 boundedness of Cauchy integrals on Lipschitz curves. *J. Amer. Math. Soc.* **2** (3) (1989), 553–564.

[CMcM] Coifman, R., McIntosh, A., and Meyer, Y., L'intégrale de Cauchy définit un opérateur borné sur L^2 pour les courbes lipschitziennes. *Ann. of Math.* **116** (1982), 361–387.

[CM] Coifman, R., and Meyer, Y., Non-linear harmonic analysis and PDE. In *Beijing Lectures in Harmonic Analysis* (ed. by E. M. Stein), Ann. of Math. Stud. 112, Princeton University Press, Princeton, NJ, 1986.

[D1] David, G., Unrectifiable 1-sets have vanishing analytic capacity. *Rev. Mat. Iberoamericana* **14** (2) (1998), 369–479.

[D2] David, G., Analytic capacity, Calderón-Zygmund operators, and rectifiability. *Publ. Mat.* **43** (1) (1999), 3–25.

[DJ] David, G., and Journé, J.-L., A boundedness criterion for generalized Calderón-Zygmund operators. *Ann. of Math.* **120** (1984), 371–398.

[DJS] David, G., Journé, J.-L., and Semmes, S., Opérateurs de Calderón-Zygmund, fonctions para-accrétives et interpolation. *Rev. Mat. Iberoamericana* **1** (1985), 1–56.

[FS] Fefferman, C., and Stein, E. M., H^p spaces of several variables. *Acta Math.* **129** (3–4) (1972), 137–193.

[HLMc] Hofmann, S., Lacey, M., and McIntosh, A., The solution of the Kato problem for divergence form elliptic operators with Gaussian heat kernel bounds. *Ann. of Math.* **156** (2002), 623–631.

[HMc] Hofmann, S., and McIntosh, A., The solution of the Kato problem in two dimensions. In *Proceedings of the Conference on Harmonic Analysis and PDE* (El Escorial, 2000), *Publ. Mat.* Extra Vol. (2002), 143–160.

[JK1] Jerison, D., and Kenig, C., An identity with applications to harmonic measure. *Bull. Amer. Math. Soc.* (*N.S.*) **2** (3) (1980), 447–451.

[JK2] Jerison, D., and Kenig, C., The Neumann problem on Lipschitz domains, *Bull. Amer. Math. Soc.* (*N.S.*) **4** (2) (1981), 203–207.

[JK3] Jerison, D., and Kenig, C., The Dirichlet problem in nonsmooth domains, *Ann. of Math.* (2) **113** (2) (1981), 367–382.

[K1] Kato, T., *Perturbation Theory for Linear Operators*. Grundlehren Math. Wiss. 132, Springer-Verlag, New York 1966.

[K2] Kato, T., Fractional powers of dissipative operators. *J. Math. Soc. Japan* **13** (1961), 246–274.

[Ke] Kenig, C., Harmonic analysis techniques for second order elliptic boundary value problems. In *CBMS Reg. Conf. Ser. Math.* 83, Amer. Math. Soc., Providence, RI, 1994

[KKPT] Kenig, C., Koch, H., Pipher, H. J., and Toro, T., A new approach to absolute continuity of elliptic measure, with applications to non-symmetric equations. *Adv. Math.* **153** (2) (2000), 231–298.

[KP] Kenig, C., and Pipher, J., The Neumann problem for elliptic equations with nonsmooth coefficients. *Invent. Math.* **113** (3) (1993), 447–509.

[Mc] McIntosh, A., Square roots of operators and applications to hyperbolic p.d.e.'s. In *Miniconference on Operator Theory and PDE*, Proc. Centre Math. Anal. Austral. Nat. Univ. 5, Austral. Nat. Univ., Canberra, 1984, 124–136.

[McM] McIntosh, A., and Meyer, Y., Algèbres d'opérateurs définis par des intégrales singulières. *C. R. Acad. Sci. Paris Sér. I Math.* **301** (1985), 395–397.

[MMV] Mattila, P., Melnikov, M., and Verdera, J., The Cauchy integral, analytic capacity, and uniform rectifiability. *Ann. of Math.* (2) **144** (1) (1996), 127–136.

[NTV1] Nazarov, F., Treil, S., and Volberg, A., Accretive system Tb-theorems on nonhomogeneous spaces. *Duke Math. J.* **113** (2) (2002), 259–312.

[NTV2] Nazarov, F., Treil, S., and Volberg, A., The Tb-theorem on non-homogeneous spaces. *Acta Math.* **190** (2) (2003), 151–239.

[P] Peetre, J., On convolution operators leaving $L^{p,\lambda}$ spaces invariant. *Ann. Mat. Pura Appl.* (4) **72** (1966), 295–304.

[S] Semmes, S., Square function estimates and the $T(b)$ Theorem. *Proc. Amer. Math. Soc.* **110** (3) (1990), 721–726.

[Sp] Spanne, S., Sur l'interpolation entre les espaces $\mathcal{L}_k{}^{p\Phi}$. *Ann. Scuola Norm. Sup. Pisa* (3) **20** (1966), 625–648.

[St] Stein, E. M., Singular integrals, harmonic functions, and differentiability properties of functions of several variables. In *Singular integrals* (Chicago, Ill., 1966), Proc. Sympos. Pure Math. 10, Amer. Math. Soc., Providence, R.I., 1967, 316–335.

[T] Tolsa, X., Painlevé's problem and the semiadditivity of analytic capacity. *Acta Math.* **190** (1) (2003), 105–149.

[V] Verchota, G., Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains. *J. Funct. Anal.* **59** (3) (1984), 572–611.

[Vo] A. Volberg, *Calderón-Zygmund capacities and operators on nonhomogeneous spaces*. CBMS Reg. Conf. Ser. Math. 100, Amer. Math. Soc., Providence, RI, 2003.

Department of Mathematics, University of Missouri, Columbia, MO 65211, U.S.A.
E-mail: hofmann@math.missouri.edu

Almost everywhere convergence and divergence of Fourier series

Sergey V. Konyagin*

Abstract. The aim of this expository paper is to demonstrate that there are several challenging problems concerning the behavior of trigonometric Fourier series almost everywhere.

Mathematics Subject Classification (2000). Primary 42A20; Secondary 42A24, 42B05, 42C10.

Keywords. Trigonometric Fourier series, Walsh system, partial sums, convergence, almost everywhere.

1. Introduction

We write $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ for the one-dimensional torus considered as the real line with $x+2\pi$ identified with x. Let $L(\mathbb{T})$ denote the class of all Lebesgue integrable functions $\mathbb{T} \to \mathbb{C}$. We associate with any function $f \in L(\mathbb{T})$ its Fourier series

$$f \sim \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{ikx},$$

where

$$\hat{f}(k) = \frac{1}{2\pi} \int_{\mathbb{T}} f(x) \exp(-ikx)\, dx,$$

The m-th partial sum of the trigonometric Fourier series of f is

$$S_m(f, x) = \sum_{k=-m}^{m} \hat{f}(k) \exp(ikx).$$

Unfortunately, the Fourier series of f does not necessarily converge to f. It is known from Du Bois-Reymond [8] that the Fourier series of a continuous function can unboundedly diverge at some point. Fejér [10] and Lebesgue [27] constructed other examples of such functions. Almost one century ago it was proven that a trigonometric series with coefficients tending to 0 can diverge everywhere. Lusin [30] constructed

*The author was supported by the Grant 05-01-00066 from the Russian Foundation for Basic Research.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

such a series of power type $\sum_{k=0}^{\infty} c_k e^{ikx}$, and Steinhaus [38] gave an example of everywhere divergent real series

$$\sum_{k=1}^{\infty}(a_k \cos(kx) + b_k \sin(kx))$$

such that $a_k \to 0, b_k \to 0$ as $k \to \infty$. The central problem was: can the Fourier series of an integrable function diverge almost everywhere (everywhere)? G. H. Hardy [13] proved that if it is the case the divergence is not too fast:

$$S_m(f, x) = o(\log m) \quad (m \to \infty) \quad \text{almost everywhere.} \tag{1}$$

A. N. Kolmogoroff [17] constructed his famous example of a function $f \in L(\mathbb{T})$ such that $S_m(f, x)$ diverges unboundedly almost everywhere. In another paper [19] he constructed an everywhere divergent Fourier series.

These prominent results caused two type of problems.

I. To find a large class of functions with almost everywhere convergent Fourier series.

II. To show that a Fourier series of every integrable function converges in some sense to the function.

Some versions of these problems are discussed in the paper. We will deal not only with the trigonometric system but also with the Walsh system. The Walsh system is the following system of functions defined on [0, 1):

$$w_0(x) = 1, \quad w_k(x) = \prod_{j=0}^{r} \left(\operatorname{sign} \sin 2^{j+1}\pi x\right)^{\varepsilon_j} \quad (k \in \mathbb{N}),$$

where $\varepsilon_0, \dots, \varepsilon_r$ are the digits in the representation of k in the dyadic system

$$k = \sum_{j=0}^{r} \varepsilon_j 2^j, \quad \varepsilon_j \in \{0, 1\},\ \varepsilon_r = 1. \tag{2}$$

This is a complete orthonormal functional system, and any function $f \in L[0, 1)$ has the Fourier–Walsh representation

$$f \sim \sum_{k=0}^{\infty} \hat{f}(k) w_k(x), \quad \hat{f}(k) = \int_0^1 f(x) w_k(x)\, dx.$$

The Fourier–Walsh partial sums are defined as

$$S_m(f, x) = \sum_{k=0}^{m-1} \hat{f}(k) w_k(x).$$

The theory and the history of Fourier–Walsh series can be found in [11].

The author is grateful to A. V. Rozhdestvenskii for a careful reading of the paper. Due to his remarks, some shortcomings have been corrected.

2. Convergence of the sequence of all partial sums

Let $\phi\colon [0,+\infty)\to[0,+\infty)$ be a nondecreasing function, $\phi(0)=0$. Denote

$$\phi(L)=\Big\{f\in L(\mathbb{T}):\int_{\mathbb{T}}\phi(|f(x)|)\,dx<\infty\Big\}.$$

For example, if $\phi(u)=u$, then $\phi(L)=L(\mathbb{T})$; if $\phi(u)=u^2$, then $\phi(L)=L^2(\mathbb{T})$.

Problem 2.1. For what functions ϕ the trigonometric Fourier series of any function $f\in\phi(L)$ converges almost everywhere?

In particular, it was not clear whether the Fourier series of a continuous function can diverge everywhere. Carleson [6] showed that if $f\in L^2(\mathbb{T})$ (i.e., f is measurable and $\int_{\mathbb{T}}|f^2(x)|dx<\infty$), then $S_m(f,x)\to f(x)$ as $m\to\infty$ almost everywhere. The condition $f\in L^2(\mathbb{T})$ in the Carleson theorem was weakened by Hunt [15], Sjölin [36]. Antonov [2] proved that if f is a measurable function and

$$\int_{\mathbb{T}}|f(x)|\log^+|f(x)|\log^+\log^+\log^+|f(x)|\,dx<\infty,$$

then $S_m(f,x)\to f(x)$ as $m\to\infty$ for almost all $x\in\mathbb{T}$. (We denote $\log^+u=\log u$ if $u\ge 1$ and $\log^+u=0$ if $u<1$.) The results in [36] were proven for Fourier–Walsh series, but the proofs can be rewritten for trigonometric series as well. The analog of Antonov's result for Fourier–Walsh series was established by Sjölin and Soria [37]. Arias-de-Reyna [1] constructed a rearrangement invariant space QA of functions with almost everywhere convergent Fourier series such that QA strictly contains Antonov's space $L\log^+L\log\log\log^+L$.

On the other hand, Körner [26] proved that if

$$\phi(u)=o(u\log\log u)\quad(u\to\infty)\tag{3}$$

then there is a function $f\in\phi(L)$ such that $\limsup_{m\to\infty}|S_m(f,x)|=\infty$ for all $x\in\mathbb{T}$. This result was improved by the author [22], [23].

Theorem 2.2. *If*

$$\phi(u)=o(u\sqrt{\log u/\log\log u})\quad(u\to\infty)$$

then there is a function $f\in\phi(L)$ such that

$$\limsup_{m\to\infty}S_m(f,x)=\infty\quad\textit{for all } x\in\mathbb{T}.$$

Also, in [22] and [23] it was proven the existence of a function $f\in L(\mathbb{T})$ satisfying

$$S_m(f,x)>\psi(m)\quad\text{for all } x \text{ and infinitely many } m$$

provided that

$$\psi(m)=o(\sqrt{\log m/\log\log m})\quad(m\to\infty).$$

This improved the result of Chen [7] asserting the existence of such a function f under a stronger supposition

$$\psi(m) = o(\log\log m) \quad (m \to \infty).$$

However, it is still unknown whether Hardy's inequality (1) can be improved. More precisely: does there exist a function $\psi(u) = o(u)$ as $u \to \infty$ such that for any $f \in L(\mathbb{T})$ the inequality

$$S_m(f, x) = o(\psi(\log m)) \quad (m \to \infty)$$

holds everywhere?

Conjecture 2.3. For any $f \in L(\mathbb{T})$

$$S_m(f, x) = o(\sqrt{\log m}) \quad (m \to \infty) \quad \text{almost everywhere.}$$

Due to Bochkarev, we know the existence of a function with almost everywhere divergent Fourier–Walsh series from a smaller functional class $\phi(L)$ than in the case of trigonometric Fourier series. In fact, the following result was proven in [5].

Theorem 2.4. *If*

$$\phi(u) = o(u\sqrt{\log u}) \quad (u \to \infty)$$

then there is a function $f \in L[0, 1)$ such that

$$\int_0^1 \phi(|f(x)|)dx < \infty$$

and the Fourier–Walsh partial sums of f unboundedly diverge almost everywhere.

Antonov extended his result to convergence of cubic partial sums for multiple trigonometric Fourier series ([3], [4]; see also [37]). An extension of Körner's result to multiple trigonometric Fourier series was made in [21].

3. Convergence of subsequences of the sequence of partial sums

Gosselin [12] proved that for any increasing sequence $\{m_j\}$ there is $f \in L(\mathbb{T})$ such that

$$\sup_j |S_{m_j}(f, x)| = \infty \tag{4}$$

for almost all $x \in \mathbb{T}$. Totik [39] established the existence of f such that (4) holds for all $x \in \mathbb{T}$.

The problem is to determine under which conditions on a sequence $\{m_j\}$ and a function ϕ the partial sums $S_{m_j}(f)$ converge to f almost everywhere for any function $f \in \phi(L)$. In particular, is it true that for enough sparse sequence $\{m_j\}$ we can claim

the almost everywhere convergence of $S_{m_j}(f)$ to f for a wider functional class $\phi(L)$ than in the case of taking the full sequence of the partial sums?

Let $\{m_j\}$ be a lacunary sequence, that is, $\inf_j m_{j+1}/m_j > 1$. If the Fourier series of a function g is a power-type series, namely,

$$g \sim \sum_{k=0}^{\infty} \hat{g}(k) \exp(ikx),$$

then the sequence $\{S_{m_j}(g)\}$ converges to g almost everywhere ([42], Chapter 15, Theorem 5.11). Combining the last result with Theorem 5.11 of Chapter 7 from [42] we get that if a measurable function f satisfies the condition

$$\int_{\mathbb{T}} |f(x)| \log^+ |f(x)| \, dx < \infty, \tag{5}$$

and $\{m_j\}$ is a lacunary sequence, then $\{S_{m_j}(f)\}$ converges to f almost everywhere. Let us recall that we do not know whether (5) is sufficient for the almost everywhere convergence of the Fourier series of the function f.

The following theorem [25] contains the above-mentioned results of [26] and [39].

Theorem 3.1. *For any increasing sequence $\{m_j\}$ of positive integers and any nondecreasing function $\phi\colon [0, +\infty) \to [0, +\infty)$ satisfying condition* (3)*, there is a function $f \in \phi(L)$ such that*

$$\sup_j |S_{m_j}(f, x)| = \infty \quad \textit{for all } x \in \mathbb{T}.$$

For the proof the construction suggested by Heladze [14] is used.

However, the technique of [22] and [23] employed the partial sums S_{m_j} for a quite rich subsequence $\{m_j\}$ and did not allow to weaken condition (3) in Theorem 3.1. Moreover, it is quite possible that this condition is sharp.

Conjecture 3.2. For any lacunary increasing sequence $\{m_j\}$ of positive integers and any measurable function f such that

$$\int_{\mathbb{T}} |f(x)| \log^+ \log^+ |f(x)| \, dx < \infty,$$

we have $S_{m_j}(f, x) \to f(x)$ as $j \to \infty$ for almost all $x \in \mathbb{T}$.

There is no a direct analog of Gosselin's theorem for Fourier–Walsh series. For a positive integer k denote

$$s(k) = 1 + \sum_{j=0}^{r-1} |\varepsilon_j - \varepsilon_{j+1}|,$$

where $\varepsilon_0, \ldots, \varepsilon_r$ are defined by (2). It is well-known that if $\{m_j\}$ is an increasing sequence of positive integers and

$$\sup_j s(m_j) < \infty, \tag{6}$$

then the Fourier–Walsh sums $S_{m_j}(f)$ converge to f almost everywhere for any $f \in L[0, 1)$. In particular, this holds for $m_j = 2^j$. Moreover, condition (6) is not necessary for almost everywhere convergence of $S_{m_j}(f)$ to f for all functions $f \in L[0, 1)$ [20].

Problem 3.3. Find a necessary and sufficient condition on a sequence $\{m_j\}$ of positive integers under which the partial Fourier–Walsh sums $S_{m_j}(f)$ converge to f almost everywhere for every function $f \in L[0, 1)$.

One of the most remarkable results in this subject belongs to Lukomskij [29] who settled the problem for multiple Fourier–Walsh series of any dimension greater than one [29].

4. Ul'yanov's problem

Kolmogoroff [18] established the weak-type estimate for conjugate functions. One of the corollaries of his result is the convergence of trigonometric Fourier series of any integrable function in $L^p (0 < p < 1)$ and, therefore, in measure. Hence, for any function $f \in L(\mathbb{T})$ there exists an increasing sequence $\{m_j\}$ of positive integers such that the partial sums $S_{m_j}(f)$ converge to f almost everywhere. Gosselin's theorem demonstrates that such a sequence $\{m_j\}$ must depend on a function f.

Ul'yanov [40] asked the following question.

Problem 4.1. Does there exist a sequence $\{M_j\}$ such that the Fourier series of any $f \in L(\mathbb{T})$ possesses a subsequence $\{S_{m_j}(f)\}$ of its partial sums converging almost everywhere to f such that $m_j \le M_j$ for all j?

The following results have been proven in [24].

Theorem 4.2. *If $\phi(u)/u \to \infty$ as $u \to \infty$ then there exists a sequence $\{M_j\}$ such that for every function $f \in \phi(L)$ there is an increasing sequence $\{m_j\}$ such that $m_j \le M_j$ for all j and $S_{m_j}(f) \to f$ almost everywhere.*

Theorem 4.3. *There exists a sequence $\{M_j\}$ such that for every function $f \in L(\mathbb{T})$ there is an increasing sequence $\{m_j\}$ such that $m_j \le M_j$ for infinitely many j and $S_{m_j}(f) \to f$ almost everywhere.*

The sequence $\{M_j\}$ from Theorem 4.3 must grow very rapidly, faster than any multiple iteration of the exponent. Define the function $\exp^*(k)$, $k \in \mathbb{N}$, as the following: $\exp^*(0) = 0$, $\exp^*(k) = e^{\exp^*(k-1)}$ $(k \ge 1)$. It turns out that for any $\varepsilon > 0$ any sequence $\{M_j\}$ with

$$M_j = O\left(\exp^*([(\log\log j)^{1-\varepsilon}])\right) \quad (j \ge 20)$$

does not satisfy the requirements of Theorem 4.3. (Here $[u]$ is the greatest integer not exceeding u.)

Now we consider Ul'yanov's problem for measures. Let μ be a Borel measure on $\mathbb{T}$ with $\int_{\mathbb{T}} |\,d\mu| < \infty$. Denote

$$\hat{\mu}(k) = \frac{1}{2\pi} \int_{\mathbb{T}} \exp(-ikx) d\mu(x), \quad S_m(\mu, x) = \sum_{k=-m}^{m} \hat{\mu}(k) \exp(ikx).$$

In general, one cannot find an increasing sequence $\{m_j\}$ such that for almost all $x \in \mathbb{T}$ a sequence $\{S_{m_j}(\mu, x)\}$ converges, but, as follows from [18], the latter sequence is bounded for some $\{m_j\}$ and for almost all $x \in \mathbb{T}$. Ul'yanov's problem for measures can be formulated as follows.

Problem 4.4. Does there exist a sequence $\{M_j\}$ such that for any measure μ there is an increasing sequence $\{m_j\}$ such that $m_j \le M_j$ and for almost all $x \in \mathbb{T}$ the sequence $\{S_{m_j}(\mu, x)\}$ is bounded?

Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence of points in $\mathbb{T}$ and $\{a_k\}_{k\in\mathbb{N}}$ be a sequence of positive numbers satisfying the conditions $a_1 \ge a_2 \ge \cdots$, and $\sum_k a_k < \infty$. Let $\sigma = \{\sigma_k\}_{k\in\mathbb{N}}$ be a sequence of independent (real or complex) uniformly bounded random variables with the zero expectations. We define the random measure $\mu = \sum_k \sigma_k a_k \delta_{x_k}$, where δ_x is the Dirac unit point mass at x. The following assertion has been proved by the author and F. L. Nazarov.

Theorem 4.5. *Let $\eta > 0$ and $M_{\eta,j} = M_j = \exp^*[\eta \log\log(j+2))]$ for $j \in \mathbb{N}$. Then for any sequences $\{x_k\}$ and $\{a_k\}$ there exists an increasing sequence $\{m_j\}$ such that $m_j \le M_j$ for all sufficiently large j and, moreover, for any σ, almost all $\omega \in \Omega$ the inequality $\sup_j |S_{m_j}(d\mu(\omega))| < \infty$ holds almost everywhere on $\mathbb{T}$.*

I can prove that in general the estimate for the growth of $\{M_j\}$ in the theorem is sharp.

5. Strong summability

There are several ways to reconstruct the values of an integrable function via its partial Fourier sums almost everywhere. By the classical theorem of Lebesgue [28], the Fejér means $\frac{1}{M+1}\sum_{m=0}^{M} S_m(f)$ converge to f almost everywhere for every integrable f. This fact can be rewritten as

$$\lim_{M\to\infty} \frac{1}{M+1} \sum_{m=0}^{M} (S_m(f,x) - f(x)) = 0$$

almost everywhere. Marcienkiewicz [31] discovered that the convergence is not connected with oscillation of positive and negative terms, but reflects the fact that for most values m the difference $S_m(f,x) - f(x)$ is small. He proved that

$$\lim_{M\to\infty} \frac{1}{M+1} \sum_{m=0}^{M} |S_m(f,x) - f(x)|^2 = 0$$

holds almost everywhere and thus created the theory of strong summability. Let $\phi\colon [0,+\infty) \to [0,+\infty)$ be a nondecreasing function, $\phi(0) = 0$. There was a problem: for which ϕ we have

$$\lim_{M\to\infty} \frac{1}{M+1} \sum_{m=0}^{M} \phi(|S_m(f,x) - f(x)|) = 0 \tag{7}$$

almost everywhere for all $f \in L(\mathbb{T})$? As faster the growth of ϕ, as stronger the result. Thus, Marcienkiewicz [31] proved (7) for $\phi(u) = u^2$, and Zygmund [41] extended it to $\phi(u) = u^p$ for any $p > 0$.

If, moreover, $\phi(u) > 0$ for $u > 0$, then (7) implies the following property of the sequence $\{s_m\} = \{S_m(f,x)\}$ and the number $s = f(x)$: there exists an increasing sequence $\{m_j\}$ such that $\lim_j m_j/j = 1$ and $\lim_{j\to\infty} s_{m_j} = s$. This property is called almost everywhere ([42], Chapter 13, (9.1)), or statistical, convergence. There are many recent papers on statistical convergence; see, e.g., [32] and [33].

So, for almost all $x \in \mathbb{T}$ there exists an increasing sequence $\{m_j\}$ with $m_j = O(j)$ such that $\lim_{j\to\infty} S_{m_j}(f,x) = f(x)$. Recall that if we are looking for the same $\{m_j\}$ for almost all $x \in \mathbb{T}$ then in general its growth must be much faster.

To describe the contemporary approach to the strong summability, we recall the notion of the bounded mean oscillation (BMO); see, e.g., ([9], Chapter 6, §2). For a nondegenerate interval $I \subset \mathbb{R}$ and a function $f \in L(I)$ denote $f_I = |I|^{-1} \int_I f(x)dx$. The set BMO$[0,\infty)$ is defined as the class of all locally integrable on $[0,\infty)$ functions f such that

$$\|f\|_* = \sup_{I\subset[0,\infty)} \left\{ |I|^{-1} \int_I |f(x) - f_I|\, dx \right\} < \infty.$$

Next, for a function $f \in L(\mathbb{T})$, $x \in \mathbb{T}$, $t \in [0,\infty)$ denote

$$g_x(t) = S_{[t]}(f,x) - f(x).$$

Let $\mu(E)$ be the Lebesgue measure of a set E. The following result was proven by Rodin in [34], [35].

Theorem 5.1. *For any function $f \in L(\mathbb{T})$ and for almost all $x \in \mathbb{T}$ we have $g_x \in$ BMO$[0,\infty)$. Moreover, there exists an absolute constant $C > 0$ such that for any $\alpha > 0$ the following inequality holds*

$$\mu\left\{ x \in \mathbb{T} :\ \|g_x\|_* > \alpha \int_{\mathbb{T}} |f(y)|\, dy \right\} \le C\alpha^{-1}.$$

Corollary 5.2. *Let $\lambda > 0$ and $\phi(u) = \exp(\lambda u) - 1$. Then equality* (7) *holds almost everywhere for all $f \in L(\mathbb{T})$.*

Corollary 5.2 easily follows from Theorem 5.1 and the John–Nirenberg inequality ([9], Chapter 6, §4).

The condition on the function ϕ in Corollary 5.2 is sharp. Karagulyan [16] proved that if

$$\limsup_{u\to\infty} \log \phi(u)/u = \infty$$

then (7) fails for some $f \in L(\mathbb{T})$ for all $x \in \mathbb{T}$.

References

[1] Arias-de-Reyna, J., Pointwise convergence of Fourier series. *J. London Math. Soc.* **65** (2002), 139–153.

[2] Antonov, N. Yu., Convergence of Fourier series. *East J. Approx.* **2** (1996), 187–196.

[3] Antonov, N. Yu., Behavior of partial sums of trigonometric Fourier series. Ph. D. Thesis, Ekaterinburg, 1998 (in Russian).

[4] Antonov, N. Yu., On the convergence almost everywhere of multiple trigonometric Fourier series over cubes. *Izv. Math.* **68** (2004), 223–241.

[5] Bochkarev, S. V., On the problem of the smoothness of functions whose Fourier–Walsh series diverge almost everywhere. *Dokl. Math.* **61** (2000), 263–266.

[6] Carleson, L., On convergence and growth of partial sums of Fourier series. *Acta Math.* **116** (1966), 133–157.

[7] Chen, Y. M., An almost everywhere divergent Fourier series of the class $L(\log^+ \log^+ L)^{1-\varepsilon}$. *J. London Math. Soc.* **44** (1969), 643–654.

[8] Du Bois-Reymond, P., Untersuchungen über die Convergenz und Divergenz der Fourierschen Darstellungsformen. *Abhand. Akad. München* **12** (1876), 1–103.

[9] Duoandikoetxea, Javier, *Fourier Analysis*. Grad. Stud. Math. 29, Amer. Math. Soc., Providence, RI, 2001.

[10] Fejér, L., Sur les singularités des séries de Fourier de fonctions continues. *Ann. Sci. École Norm. Sup.* **28** (1911), 63–103.

[11] Golubov, B., Efimov, A., Skvortsov, V., *Walsh series and transforms. Theory and applications*. Math. Appl. (Soviet Ser.) 64, Kluwer Academic Publishers, Dordrecht 1991.

[12] Gosselin, R. P., On the divergence of Fourier series. *Proc. Amer. Math. Soc.* **9** (1958), 278–282.

[13] Hardy, G. H., On the summability of Fourier's series. *Proc. London Math. Soc.* **12** (1913), 365–372.

[14] Heladze, Š. V., On the everywhere divergence of Fourier series of functions from a class $L\phi(L)$. *Trudy Tbiliss. Mat. Inst.* **89** (1988), 51–59 (in Russian).

[15] Hunt, R. A., On the convergence of Fourier series. In *Orthogonal expansions and their continuous analogues*, Southern Ill. University Press, Carbondale, Ill., 1968, 235–255.

[16] Karagulyan, G. A., On the divergence of strong Φ-means of Fourier series. *J. Contemp. Math. Anal.* **26** (1991), 66–69.

[17] Kolmogoroff, A. N., Une série de Fourier–Lebesgue divergente presque partout. *Fund. Math.* **4** (1923), 324–328.

[18] Kolmogoroff, A. N., Sur les fonctions harmoniques conjuguées et les séries de Fourier. *Fund. Math.* **7** (1925), 23–28.

[19] Kolmogoroff, A. N., Une série de Fourier–Lebesgue divergente partout. *C. R. Acad. Sci. Paris* **183** (1926), 1327–1329.

[20] Konyagin, S. V., The Fourier-Walsh subsequence of partial sums. *Math. Notes* **54** (1993), 1026–1030.

[21] Konyagin, S. V., On divergence of trigonometric Fourier series over cubes. *Acta Sci Math.* (*Szeged*) **61** (1995), 305–329.

[22] Konyagin, S. V., On divergence of trigonometric Fourier series everywhere. *C. R. Acad. Sci. Paris Sér. I Math.* **329** (1999), 693–697.

[23] Konyagin, S. V., On the divergence everywhere of trigonometric Fourier series. *Sb. Math.* **191** (2000), 97–120.

[24] Konyagin, S. V., Convergent subsequences of partial sums of Fourier series of $\varphi(L)$. In *Orlicz centenary volume*, Banach Center Publ. 64, Institute of Mathematics, Polish Academy of Sciences, Warsaw 2004, 117–126.

[25] Konyagin, S. V., Divergence everywhere of subsequences of partial sums of trigonometric Fourier series. *Proc. Steklov Inst. Math.* Suppl. 2 (2005), 167–175.

[26] Körner, T. W., Everywhere divergent Fourier series. *Colloq. Math.* **45** (1981), 103–118.

[27] Lebesgue, H., Sur les intégrales singulières. *Ann. Fac. Sci. Toulouse* **1** (1909), 25–117.

[28] Lebesgue, H., Sur la représentation trigonométrique approchée des fonctions à une conditiom de Lipschitz. *Bull. Soc. Math. France* **38** (1910), 184–210.

[29] Lukomskij, S. F., Multiple Walsh series: Convergence in measure and almost everywhere. *Dokl. Math.* **57** (1998), 81–82.

[30] Lusin, N. N., Über eine Potenzreihe. *Rend. Circ. Mat. Palermo* **32** (1911), 386–390.

[31] Marcienkiewicz, J., Sur l'interpolation. *Stud. Math.* **6** (1936), 1–17, 67–81.

[32] Móricz, F., Statistical convergence of Walsh–Fourier series. *Acta Math. Acad. Paedagog. Nyházi.* **20** (2004), 165–168.

[33] Móricz, F., Regular statistical convergence of double sequences. *Colloq. Math.* **102** (2005), 217–227.

[34] Rodin, V. A., BMO–strong means of Fourier series. *Funct. Anal. Appl.* **23** (1989), 145–147.

[35] Rodin, V. A., The space BMO and strong means of Fourier series. *Anal. Math.* **16** (1990), 291–302.

[36] Sjölin, P., An inequality of Paley and convergence a.eòf Walsh–Fourier series. *Ark. Mat.* **7** (1969), 551–570.

[37] Sjölin, P., Soria, F., Remarks on a theorem by N. Yu. Antonov. *Stud. Math.* **158** (2003), 79–97.

[38] Steinhaus, H., Une série trigonométrique partout divergente. *Compt. Rend. Soc. Scient. de Varsovie* (1912), 219–229.

[39] Totik, V., On the divergence of Fourier series. *Publ. Math.* (*Debrecen*) **29** (1982), 251–264.

[40] Ul'yanov, P. L., Solved and unsolved problems in the theory of trigonometric and orthogonal series. *Russian Math. Surveys* **19** (1964), 1–62.

[41] Zygmund, A., On the convergence and summability of power series on the circle of convergence (II). *Proc. London Math. Soc.* **47** (1941), 326–350.

[42] Zygmund, A., *Trigonometric series*. Vol. 1, 2, Cambridge University Press, Cambridge, London, New York, Melbourne 1977.

Department of Mechanics and Mathematics, Moscow State University, 119992 Moscow, Russia

E-mail: konyagin@ok.ru

Iterated Segre mappings of real submanifolds in complex space and applications

Linda Preiss Rothschild*

Abstract. This article is a survey of various applications of the method of iterated Segre mappings obtained by a number of mathematicians, including the author, over the past decade. This method is applied to various problems involving real submanifolds in complex space and their mappings. The article begins with a description of the iterated Segre mappings associated to generic submanifolds. The problems addressed concern transversality of holomorphic mappings, finite jet determination, local stability groups, and algebraicity of holomorphic mappings between real-algebraic manifolds.

Mathematics Subject Classification (2000). Primary 32H02, 32H35, 32V40; Secondary 14P20.

Keywords. Generic submanifolds, holomorphic mappings, Segre mappings, finite determination.

1. Introduction and notation

In this survey we consider real submanifolds $M \subset \mathbb{C}^N$ and $M' \subset \mathbb{C}^{N'}$ through p and p' respectively and study local properties of germs of holomorphic mappings $H\colon (\mathbb{C}^N, p) \to (C^{N'}, p')$ such that $H(M) \subset M'$. The questions we consider are the following:

1. When is H determined by finitely many derivatives at p?
2. If $M = M'$ and $p = p'$, when does the set of all such H form a finite dimensional Lie group?
3. If M and M' are real-algebraic manifolds, when does it follow that H is necessarily an algebraic mapping?
4. When is H transversal to M' at p'?

The common thread of the approach to these problems is the use of the iterated Segre mappings, which represent a kind of blow-up of the complexification of a real submanifold. These mappings first appeared in the joint work of the author with

*The author is partially supported by National Science Foundation grant DMS-0400880.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Baouendi and Ebenfelt [2]. In what follows I shall assume, without loss of generality, that p and p' are both the origin. Then the real submanifold M is given near 0 by the vanishing of a system of equations $\rho = (\rho_1, \ldots, \rho_d) = 0$, where d is the codimension of M and the ρ_j are real-valued functions with linearly independent differentials at 0. If, in addition, the complex vectors $\left(\frac{\partial \rho_j}{\partial Z_1}, \ldots, \frac{\partial \rho_j}{\partial Z_N}\right)(0)$, $1 \leq j \leq d$, are also linearly independent, the real submanifold M is called *generic*. In particular, the condition that M is generic insures that M is a *CR manifold*, i.e. that the space $T_q^{0,1}M$ of complex $(0, 1)$ vectors at q that are tangent to M at q is of constant dimension for $q \in M$ near 0. The assumption that a CR submanifold of $\mathbb{C}^N$ is generic is not very restrictive, since any such manifold can be locally embedded as a generic submanifold in some complex manifold, possibly of lower dimension (see e.g. [6]).

For quite some time an important tool in the study of holomorphic mappings has been the complexification of the real manifolds M and M' (see e.g. Webster [48], [49], Diederich-Webster [25], Baouendi–Jacobowitz–Treves [12], Diederich–Pinchuk [24], Chern–Ji [19]). To simplify notation and statements, I shall assume, unless stated otherwise, that the real submanifolds M and M' are real-analytic, and hence their defining functions can be assumed to be real-analytic. Then we may regard the real vector valued function ρ as a convergent power series, $\rho(Z, \bar{Z})$, in Z and $\bar{Z}$, which may be complexified as a (germ at 0 of a) convergent power series $\rho(Z, \zeta)$ in $2N$ complex variables. The *complexification* $\mathcal{M}$ of M is defined to be the (germ at 0 of a) complex manifold $\mathcal{M}$ defined near $(0, 0) \in \mathbb{C}^N \times \mathbb{C}^N$ by

$$\mathcal{M} = \{(Z, \zeta) \in \mathbb{C}^N \times \mathbb{C}^N : \rho(Z, \zeta) = 0\}. \tag{1.1}$$

(Here the real-analyticity of M and ρ allows the local complexification of ρ. By abuse of notation, I shall denote both ρ and its local complexification by the same letter.) Similarly, I denote by $\mathcal{M}'$ the complexification of M'. For $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^{N'}, 0)$ denote by $\mathcal{H} = (H, \overline{H})\colon (\mathbb{C}^N \times \mathbb{C}^N, 0) \to (\mathbb{C}^{N'} \times \mathbb{C}^{N'}, 0)$ the *complexification* of H. (Here $\overline{H}(\zeta) := \overline{H(\bar{\zeta})}$.) Using again real-analyticity, it is easy to see that

$$H(M) \subset M' \iff \mathcal{H}(\mathcal{M}) \subset \mathcal{M}'. \tag{1.2}$$

The submanifold M is said to be of *finite type* at 0 in the sense of Kohn [40] (and also of Bloom–Graham) if the (complex) Lie algebra $\mathfrak{g}_M$ generated by all smooth $(1, 0)$ and $(0, 1)$ vector fields tangent to M satisfies $\mathfrak{g}_M(0) = \mathbb{C}T_0M$, where $\mathbb{C}T_0M$ is the complex tangent space to M at 0. The method of Segre mappings is an important tool in the analysis of the equation on the right hand side of (1.2) in case M and M' are of finite type at the origin.

In this survey I shall focus mainly on results for the case $N = N'$. Also, for simplicity of notation, the (germ at 0 of a) holomorphic map $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ will be assumed to be finite, i.e. any $Z \in \mathbb{C}^N$ near 0 has only finitely many inverse images under H near 0. An equivalent algebraic condition is that the ideal $I(H)$ in $\mathbb{C}\{Z\}$ (the ring of convergent power series in $Z = (Z_1, \ldots, Z_N)$) generated by the

components, $H_1(Z), \dots, H_N(Z)$, of $H(Z)$ is of finite codimension, i.e. the vector space $\mathbb{C}\{Z\}/I(H)$ is finite dimensional.

2. Iterated Segre mappings

Following the approach of [8], I shall give here an outline of the definition and main properties of the iterated Segre mappings for a generic submanifold M of codimension d in $\mathbb{C}^N$. Let $n := N - d$ and $\gamma\colon (\mathbb{C}^N \times \mathbb{C}^n, 0) \to (\mathbb{C}^N, 0)$ be a (germ of a) holomorphic mapping such that

$$\rho(\gamma(\zeta, t), \zeta) \equiv 0, \quad \operatorname{rk} \frac{\partial \gamma}{\partial t}(0,0) = n, \tag{2.1}$$

where $\zeta = (\zeta_1, \dots, \zeta_N)$, $t = (t_1, \dots, t_n)$. By the implicit function theorem, the existence of such γ follows from the genericity assumption on M. We define a sequence of germs of holomorphic mappings $v^j\colon (\mathbb{C}^{nj}, 0) \to (\mathbb{C}^N, 0)$, $j \geq 0$, called *the iterated Segre mappings* of M at 0 (relative to γ), inductively as follows:

$$\begin{aligned} &v^0 = 0, \\ &v^1(t^1) := \gamma(0, t^1), \\ &v^{j+1}(t^1, \dots, t^{j+1}) := \gamma\big(\overline{v^j}(t^1, \dots, t^j), t^{j+1}\big). \end{aligned} \tag{2.2}$$

Here $\overline{v^j}$ denotes the vector-valued convergent power series obtained from v^j by conjugating its coefficients.

For a (germ of a) holomorphic mapping $v\colon (\mathbb{C}^k, 0) \to (\mathbb{C}^l, 0)$, we denote by $\operatorname{Rk} v$ the generic rank of any representative of v in a sufficiently small neighborhood of 0. Some of the main properties of the Segre mappings are summarized in the following theorem.

Theorem 2.1 ([8]). *Let $M \subset \mathbb{C}^N$ be a real-analytic generic submanifold through 0 of codimension d, and let γ, v^j be as above. Then:*

(i) *For $l = 1, 2, \dots$, the mapping $(v^l(t^1, \dots, t^l), \overline{v^{l-1}}(t^1, \dots, t^{l-1}))$ sends $\mathbb{C}^{nl}$ into $\mathcal{M}$.*

(ii) *There exists an integer k_0, $1 \leq k_0 \leq d+1$, such that $\operatorname{Rk} v^j = \operatorname{Rk} v^{j+1}$ for $j \geq k_0$, and if $k_0 > 1$, then $\operatorname{Rk} v^j < \operatorname{Rk} v^{j+1}$ for $1 \leq j \leq k_0 - 1$.*

If k_0 is as in (ii)*, the following hold.*

(iii) *The submanifold M is of finite type at 0 if and only if $\operatorname{Rk} v^{k_0} = N$.*

(iv) *There exists a (germ at 0 of a) submanifold $\Sigma \subset \mathbb{C}^{2nk_0}$, such that $v^{2k_0}(\Sigma) = \{0\}$ and v^{2k_0} achieves full rank (i.e. $\operatorname{Rk} v^{k_0}$) on some points on Σ (arbitrarily close to 0).*

Here are some observations about Theorem 2.1. Part (ii) shows that the generic ranks of the Segre mappings stabilize before $d+1$ iterations. Part (iii) gives a characterization of finite type in terms of the generic rank of the Segre mappings after stabilization. Part (iv) shows that after at most $2(d+1)$ iterations, the Segre mappings achieve maximal rank on a set that is mapped by v^{2k_0} to the origin in $\mathbb{C}^N$.

The Segre mappings can be used in conjunction with special choices of coordinates in $\mathbb{C}^N$ for a given generic submanifold M. The coordinates $(z, w) \in \mathbb{C}^N = \mathbb{C}^n \times \mathbb{C}^d$ are called *normal* if there exists a (germ of a) holomorphic mapping $Q(z, \chi, \tau)$, $Q\colon (\mathbb{C}^n \times \mathbb{C}^n \times \mathbb{C}^d, 0) \to (\mathbb{C}^d, 0)$, with $Q(z, 0, \tau) \equiv Q(0, \chi, \tau) \equiv \tau$, such that M is given near the origin by the vector equation

$$w = Q(z, \overline{z}, \overline{w}). \tag{2.3}$$

The existence of normal coordinates may be proved by the use of the implicit function theorem ([20], [6]). If $Z = (z, w)$ is a choice of normal coordinates for M given by (2.3), then the equation in (2.1) for $\gamma(\zeta, t) = (\gamma_1(\zeta, t), \gamma_2(\zeta, t)) \in \mathbb{C}^n \times \mathbb{C}^d$ becomes

$$\gamma_2(\zeta, t) \equiv Q(\gamma_1(\zeta, t), \zeta). \tag{2.4}$$

Hence we may choose

$$\gamma(\zeta, t) = ((\gamma_1(\zeta, t), \gamma_2(\zeta, t)) = (t, Q(t, \zeta)). \tag{2.5}$$

Now suppose $(z', w') \in \mathbb{C}^{N'} = (\mathbb{C}^{n'} \times \mathbb{C}^{d'})$ is a given set of normal coordinates for a generic submanifold $M' \subset \mathbb{C}^{N'}$ of codimension d'. Then if $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^{N'}, 0)$ we may write

$$H(Z) = (F(Z), G(Z)), \quad \text{where } z' = F(Z),\ w' = G(Z). \tag{2.6}$$

By (1.2), $H(M) \subset M'$ is equivalent to the holomorphic vector equation

$$G(Z) = Q'(F(Z), \overline{F}(\zeta), \overline{G}(\zeta)) \quad \text{for all } (Z, \zeta) \in \mathcal{M}. \tag{2.7}$$

By (i) of Theorem 2.1, if (2.7) holds, then for any integer $k \geq 0$ we may take $(Z, \zeta) = (v^{k+1}(t^1, \ldots, t^{k+1}), \overline{v^k}(t^1, \ldots, t^k))$ in equation (2.7) to obtain

$$G \circ v^{k+1} = Q'\big(F \circ v^{k+1}, \overline{F \circ v^k}, \overline{G \circ v^k}\big), \tag{2.8}$$

yielding a holomorphic vector equation in $n(k+1)$ free complex parameters. If M is of finite type at 0 and (2.8) holds for $k \geq d+1$, then by Theorem 2.1 (iii) it follows that (2.7) also holds. Hence (2.8) is equivalent to (2.7) in this case.

A construction analogous to that of the iterated Segre mappings was developed independently by Christ–Nagel–Stein–Wainger [21] in a different context.

3. Nondegeneracy conditions for generic submanifolds

In order to formulate nontrivial results concerning the questions posed in Section 1, we recall some nondegeneracy conditions that may be imposed on generic submanifolds. The best known (and strongest) of these conditions is that of Levi-nondegeneracy for a real hypersurface in $\mathbb{C}^N$. If M is a hypersurface given in normal coordinates by the (scalar) equation (2.3), then M is *Levi-nondegenerate* at 0 if the quadratic form

$$\mathbb{C}^{N-1} \times \mathbb{C}^{N-1} \ni (z, \chi) \mapsto Q^{(2)}(z, \chi) \tag{3.1}$$

is nondegenerate, where $Q^{(2)}(z, \chi)$ is the quadratic part of the Taylor expansion at 0 of $Q(z, \chi, 0)$. (Equivalent definitions independent of the choice of coordinates can be given for Levi-nondegeneracy, as well as for the other nondegeneracy conditions to be discussed here.) Germs of Levi-nondegenerate hypersurfaces have been completely classified up to local equivalence in the celebrated work of Chern and Moser [20] in the 1970s.

For a generic submanifold M of any codimension, given in normal coordinates by (2.3), a number of weaker geometric conditions can be given in terms of the vector-valued function $Q(z, \chi, \tau)$. We write $Q = (Q_1, \ldots, Q_d)$. The strongest of these conditions is finite nondegeneracy: M is *ℓ-nondegenerate* at 0 if the set of vectors

$$\frac{\partial^\alpha}{\partial \chi^\alpha}\left(\frac{\partial Q_j}{\partial z_1}, \ldots, \frac{\partial Q_j}{\partial z_n}\right)(0), \quad j = 1, \ldots, d,\ \alpha \in \mathbb{N}^n,\ |\alpha| \le \ell, \tag{3.2}$$

spans $\mathbb{C}^n$, and ℓ is the smallest positive integer for which this is true. If M is ℓ-nondegenerate at 0 for some finite integer ℓ, then M is called *finitely nondegenerate*. A hypersurface is Levi-nondegenerate at 0 if and only if it is 1-nondegenerate.

A condition weaker than finite nondegeneracy is the following. Expand $Q(z, \chi, 0)$ as a Taylor series in χ,

$$Q(z, \chi, 0) = \sum Q^\alpha(z, 0, 0)\chi^\alpha, \tag{3.3}$$

and let I be the ideal in $\mathbb{C}\{z\}$ generated by all the $Q^\alpha(z, 0, 0)$. Then M is said to be *essentially finite* at 0 if the ideal I is of finite codimension in $\mathbb{C}\{z\}$. If so, the dimension of the vector space $\mathbb{C}\{z\}/I$ is called the *essential type* of M at 0, denoted $\mathrm{ess}_0 M$. It is easy to see that if M is finitely nondegenerate at 0, then it is essentially finite at 0. However, the converse is not true. For example, the hypersurface $M \subset \mathbb{C}^2$ given as $\{(z, w) \in \mathbb{C}\times\mathbb{C} : \mathrm{Im}\, w = |z|^4\}$ is essentially finite at 0, but not finitely nondegenerate. The precise relationship is the following (see [6], Proposition 11.8.27):

$$M \text{ is finitely nondegenerate at } 0 \iff M \text{ is essentially finite at } 0 \text{ with } \mathrm{ess}_0 M = 1. \tag{3.4}$$

The conditions of finite nondegeneracy and essential finiteness can be expressed invariantly, i.e. independently of any choice of defining function or coordinates (see [6], Chapter XI), but I shall not do so here. However, I want to point out that the

condition of finite type, as defined in Section 1, cannot be easily expressed in terms of an arbitrary defining function, except in the case of hypersurfaces. This is one reason that Theorem 2.1 (iii) and (iv) turns out to be useful for the study of generic submanifolds of finite type.

Another condition, which is weaker than essential finiteness, is most easily expressed in terms of $(1,0)$ vector fields, i.e. sections of the vector bundle $T^{1,0}\mathbb{C}^N$ of $(1,0)$ vectors on $\mathbb{C}^N$. The generic submanifold M is said to be *holomorphically nondegenerate* at 0 if there is no nonzero $(1,0)$ vector field with holomorphic coefficients that is tangent to M in a whole neighborhood of 0.

It can be shown (see e.g. [6], Corollary 11.7.28) that holomorphic degeneracy of a connected generic submanifold M can also be described in terms of finite nondegeneracy or essential finiteness. In fact, if M is a connected, real-analytic generic submanifold of $\mathbb{C}^N$ with $0 \in M$, then the following conditions are equivalent:

(i) M is holomorphically nondegenerate at 0.

(ii) M is holomorphically nondegenerate at all $p \in M$.

(iii) M is essentially finite at some point in M.

(iv) M is essentially finite at all points outside a proper real-analytic subset of M.

(v) M is finitely nondegenerate at some point in M.

(vi) M is finitely nondegenerate at all points outside a proper real-analytic subset of M.

(vii) There is an integer $\ell \leq N-1$ such that M is ℓ-nondegenerate at all points outside a proper real-analytic subset of M.

The relation of the above conditions with finite type is more complicated. If a hypersurface is essentially finite at 0, it is necessarily of finite type at 0, but no implication holds for generic submanifolds of higher codimension.

In view of the equivalence of (i) and (ii), if M is holomorphically nondegenerate at 0, one simply says that M is holomorphically nondegenerate. If M is not holomorphically nondegenerate, then it is not essentially finite at any point. In this sense, holomorphic nondegeneracy is the weakest condition that guarantees the existence of some "good" points.

The notion of essential finiteness has been implicitly used in the work of Diederich–Webster [25] and was defined first in the work of Baouendi–Jacobowitz–Treves [12]. Holomorphic nondegeneracy was first introduced for hypersurfaces by Stanton [47] and for higher codimension in [2]. Finite nondegeneracy was implicitly used by Han [34] for hypersurfaces and defined in joint work with Baouendi and Huang [10]. An invariant definition for CR manifolds, not necessarily embedded in complex space, is due to Ebenfelt (see [6]).

4. Transversality of mappings

The notion of transversality of real differentiable mappings has been central in differential geometry. Recall that if $f\colon (\mathbb{R}^k, 0) \to (\mathbb{R}^\ell, 0)$ is a germ of a smooth mapping, and if $E \subset \mathbb{R}^\ell$ is a smooth manifold through 0, then f is called *transversal* to E at 0 if

$$T_0 E + df(T_0\mathbb{R}^k) = T_0\mathbb{R}^\ell. \tag{4.1}$$

If $E \subset \mathbb{C}^\ell$ is a generic submanifold, and $H\colon (\mathbb{C}^k, 0) \to (\mathbb{C}^\ell, 0)$ is a (germ of a) holomorphic mapping, a stronger notion of transversality is needed, for example, to guarantee that $H^{-1}(E)$ is again a generic submanifold of $\mathbb{C}^k$. Also, for mappings in which the target is a real hypersurface, it is desirable to have a notion of transversality that guarantees the nonvanishing of a derivative of the transversal component. For generic submanifolds and holomorphic mappings as above, the appropriate notion is the following. The mapping H is said to be *CR transversal* to E at 0 if

$$T_0^{1,0} E + dH(T_0^{1,0}\mathbb{C}^k) = T_0^{1,0}\mathbb{C}^\ell. \tag{4.2}$$

Here $T_0^{1,0}E$ denotes the space of $(1, 0)$ vectors tangent to M at 0. It is not hard to see that CR transversality implies transversality, but the converse is not necessarily true. However, we shall restrict ourselves here to the case $k = \ell$ and assume that H maps one generic submanifold into another. In this context it can be shown that the two notions of transversality coincide (see [30]).

Recently Ebenfelt and the author have obtained the following result.

Theorem 4.1 ([30]). *Let $M, M' \subset \mathbb{C}^N$ be real-analytic generic submanifolds of the same dimension through 0 such that either M or M' is of finite type at 0. Then any finite holomorphic mapping $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ with $H(M) \subset M'$ is CR transversal to M' at 0.*

For the case where M and M' are assumed to be hypersurfaces, Theorem 4.1 was proved earlier by Baouendi and the author (see [14]). The case of generic submanifolds of higher codimension had remained an open problem since then.

Theorem 4.1 can be viewed as a kind of complex Hopf Lemma. For smooth mappings between hypersurfaces where the target has some convexity properties, results of this type were previously obtained by Fornaess [31], [32] for the case of pseudoconvex hypersurfaces (using the classical Hopf Lemma). Related results could also be found in the author's joint work with Baouendi [15], as well as with Baouendi and Huang [11].

I shall describe briefly here how the iterated Segre mappings can be used in the proof of Theorem 4.1. We begin with normal coordinates (z, w) and (z', w') for M and M' respectively, leading to the equation (2.7). The condition of CR transversality in these coordinates is equivalent to

$$\det \frac{\partial G}{\partial w}(0) \neq 0, \tag{4.3}$$

where $H = (F, G)$ as in (2.6). Now assume M is of finite type at 0 and let k_0 be given by Theorem 2.1 and take $k = 2k_0 - 1$ in (2.8) to obtain

$$G \circ v^{2k_0} = Q'\big(F \circ v^{2k_0}, \overline{F \circ v^{2k_0-1}}, \overline{G \circ v^{2k_0-1}}\big), \tag{4.4}$$

which is equivalent to (2.7) since M is of finite type at 0. Put $v^{2k_0}(t) = (t^{2k_0}, u^{2k_0}(t))$, with $t = (t^1, \dots, t^{2k_0}) = (t', t^{2k_0}) \in \mathbb{C}^{2nk_0-n} \times \mathbb{C}^n$. If $\Sigma \subset \mathbb{C}^{2nk_0}$ is a submanifold given by Theorem 2.1 (iv), then necessarily $\Sigma \subset \mathbb{C}^{2nk_0-n} \times \{0\}$. Taking $t^{2k_0} = 0$ in (4.4), differentiating in t', and evaluating at any $s \in \Sigma$ and using the chain rule, the left hand side of (4.4) becomes

$$\frac{\partial G}{\partial w}(0)\frac{\partial u^{2k_0}}{\partial t'}(s), \tag{4.5}$$

since $v^{2k_0} \equiv 0$ on Σ. Now if (4.3) fails, there is a nonzero constant vector $V \in \mathbb{C}^d$ such that $V^\tau \dfrac{\partial G}{\partial w}(0) = 0$. Our proof proceeds by multiplying both sides of (4.4) by V^τ, differentiating in t' and restricting to Σ. We then show that under the hypotheses of the theorem, the right hand side cannot vanish identically. The details can be found in [30].

By using formal power series, instead of convergent ones, we can also prove Theorem 4.1 when M and M' are merely assumed to be smooth and H is replaced by a CR mapping that is assumed to be "formally" finite.

One may ask also when a finite holomorphic mapping H satisfying $H(M) \subset M'$ is not only transversal to M', but actually is a diffeomorphism at 0. If M is essentially finite at 0, then we prove, using Theorem 4.1, that M' is also essentially finite at 0 and

$$\operatorname{ess}_0 M = \operatorname{mult}(H) \cdot \operatorname{ess}_0 M', \tag{4.6}$$

where ess_0 denotes the essential type at 0 and $\operatorname{mult}(H)$ denotes the multiplicity of H as a finite holomorphic mapping. If M is finitely nondegenerate at 0, then as noted in (3.4), it follows that it is essentially finite at 0 with $\operatorname{ess}_0 M = 1$. Since $\operatorname{ess}_0 M'$ is a positive integer, the equation (4.6) implies that $\operatorname{mult}(H) = 1$, i.e., H is a diffeomorphism. Hence we have the following consequence of Theorem 4.1:

Theorem 4.2 ([30]). *Let $M \subset \mathbb{C}^N$ be a real-analytic generic submanifold of finite type and finitely nondegenerate at 0. If $M' \subset \mathbb{C}^N$ is a real-analytic generic submanifold of the same dimension and $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ is a finite holomorphic mapping with $H(M) \subset M'$, then H is a local biholomorphism at 0.*

Although this paper is mostly focused on the case where the manifolds M and M' are equidimensional and contained in the same complex space, I should mention that in recent years there have been a number of related results for nonequidimensional hypersurfaces. In particular, without striving for completeness, I would like to mention the work of D'Angelo [23], Forstnerič [33], Huang [35], Huang–Ji [36], Ebenfelt–Huang–Zaitsev [27], Baouendi–Huang [9].

5. Finite jet determination

One of the most basic questions to be asked about a (germ of a) holomorphic mapping $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ is to find the least data to determine H. Clearly H is determined by $\partial^\alpha H(0)$ for all $\alpha \in \mathbb{Z}^N$. If $H(M) \subset M'$ for real analytic generic submanifolds M and M', then H also satisfies a system of real analytic equations when restricted to M, and under favorable circumstances may be determined by finitely many derivatives at the origin. We denote by $j^k_p H$ the *k-jet* of H at p, i.e.

$$j^k_p H := (\partial^\alpha H)(p)_{\{|\alpha|\le k\}}. \tag{5.1}$$

It follows from the work of Chern–Moser [20] that if H is a diffeomorphism and M and M' are both Levi-nondegenerate hypersurfaces in $\mathbb{C}^N$ through 0, then H is actually determined by $j^2_0 H$.

A striking theorem of finite determination for hypersurfaces in $\mathbb{C}^2$ is the following.

Theorem 5.1 (Ebenfelt–Lamel–Zaitsev, [29]). *Let $M \subset \mathbb{C}^2$ be a real-analytic hypersurface of finite type at 0. Then if $H^1, H^2\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ are (germs of) holomorphic diffeomorphisms both sending M into itself and satisfying $j^2_0 H^1 = j^2_0 H^2$, then $H^1 = H^2$.*

The weakest known conditions, for finite determination, that can be imposed on M and M' in higher codimenson is that of finite type and holomorphic nondegeneracy at 0. The following result is joint work with Baouendi and Mir [13].

Theorem 5.2 ([13]). *Let M and M' be real-analytic generic submanifolds of $\mathbb{C}^N$ through 0 of the same dimension, with M of finite type and holomorphically nondegenerate at 0. Let $H^0\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ be a finite holomorphic mapping with $H^0(M) \subset M'$. Then there exists an integer K such that if $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ is any finite holomorphic mapping sending M into M' with $j^K_0 H = j^K_0 H^0$, it follows that $H = H^0$.*

The techniques of the proof do not provide an explicit integer K nor give any kind of dependence of this integer on the base point. Theorem 5.2 can be strengthened by merely assuming M and M' to be smooth and by a slight weakening of the assumption that H be finite (see [13]).

For more explicit results in $\mathbb{C}^N$, $N > 2$, the condition of finite nondegeneracy might need to be imposed. This condition enters as follows. For simplicity, let us assume that $M = M'$. We wish to describe those mappings $H\colon (\mathbb{C}^N, 0) \to (\mathbb{C}^N, 0)$ that are diffeomorphisms at 0 and satisfy $H(M) \subset M$. If M is ℓ_0-nondegenerate at 0, by taking appropriate derivatives of (2.7) and applying the implicit function theorem, the following can be shown (see [6]). For $\ell \ge 0$, there exists a holomorphic mapping Ψ_ℓ defined in a neighborhood of $(0, 0, j^{\ell_0+\ell}_0 \mathrm{Id})$ in $\mathbb{C}^N \times \mathbb{C}^N \times \mathbb{C}^J$ (for an appropriate integer J) such that one has the following "basic identity". For every $(Z, \zeta) \in \mathcal{M}$

near 0 and every H sending M into itself with $j_0^{\ell_0+\ell}H$ sufficiently close to $j_0^{\ell_0+\ell}\mathrm{Id}$, the following holds:

$$(Z, j_Z^{\ell}H) = \Psi_\ell\big(Z, (\zeta, j_\zeta^{\ell_0+\ell}\overline{H})\big). \tag{5.2}$$

We now use (5.2), by taking $(Z,\zeta) = (v^k(t^1,\dots,t^k), \overline{v^{k-1}}(t^1,\dots,t^{k-1})) \in \mathcal{M}$ (as defined in Section 2). In this way, we express all derivatives of H of length $\le \ell$ on the image of v^k in terms of the derivatives of $\overline{H}$ of length $\le \ell_0+\ell$ on the image of $\overline{v^{k-1}}$. We begin by taking $(Z,\zeta) = (v^1(t^1), \overline{v^0}) = ((t^1,0),(0,0)) \in \mathcal{M}$ in (5.2) to obtain

$$\big((t^1,0), j^{\ell}_{(t^1,0)}H\big) = \Psi_\ell\big((t^1,0),((0,0), j_0^{\ell_0+\ell}\overline{H})\big). \tag{5.3}$$

Hence all derivatives of H up to order ℓ on v^1 are determined by the $\ell_0+\ell$ jet of $\overline{H}$ at 0. Taking $(Z,\zeta) = (v^2(t^1,t^2), \overline{v^1}(t^1))$, we find that all derivatives of H up to order ℓ on v^2 are determined by the $\ell_0+\ell$ jet of $\overline{H}$ on $\overline{v^1}$. Since the latter is determined by $j_0^{2\ell_0+\ell}H$ by the first step (after complex conjugation), we observe that $j^{\ell}_{v^2}H$ is determined by $j_0^{2\ell_0+\ell}H$. Now let k_0 be as in Theorem 2.1. By using inductively the above argument, we conclude that $j^0_{v^{k_0}(t)}H$ is determined by $j_0^{k_0\ell_0}H$. If M is of finite type at 0, then by Theorem 2.1 (iii), H is completely determined by the values of $j^0_{v^{k_0}(t)}H$ as t varies in $\mathbb{C}^{nk_0}$. Since $k_0 \le d+1$, the above argument gives an outline of a proof of the following.

Theorem 5.3 ([4]). *Let $M \subset \mathbb{C}^N$ be a real-analytic generic submanifold of finite type and ℓ_0-nondegenerate at 0. If $H^i\colon (\mathbb{C}^N,0) \to (\mathbb{C}^N,0)$, $i=1,2$ are holomorphic diffeomorphisms both sending M into itself and satisfying $j_0^{(d+1)\ell_0}H^1 = j_0^{(d+1)\ell_0}H^2$, it follows that $H^1 = H^2$.*

The results stated in this paper are given for (germs of) holomorphic mappings, but many can also be formulated for (germs of) smooth generic submanifolds and smooth CR mappings that are diffeomorphisms. Since the methods described above rely on unique determination by the Taylor series at 0, different techniques must be used to prove finite determination for mappings that are merely smooth. By using the method of complete systems, Ebenfelt [26] proved unique determination results for smooth CR mappings between finitely nondegenerate hypersurfaces. These were later generalized by Ebenfelt–Lamel [28]. General results for smooth CR diffeomorphisms between finitely nondegenerate smooth generic submanifolds of any codimension were recently obtained by Kim–Zaitsev [39]. The reader is referred to the survey article of Zaitsev [52] for a more detailed discussion of this topic.

6. Stability groups

In this section I shall discuss questions concerned with the structure of the group of holomorphic mappings that send a generic submanifold into itself. Let $M \subset \mathbb{C}^N$

be (the germ at 0 of) a real-analytic generic submanifold. The *holomorphic stability group*, $\mathrm{Hol}\,(M,0)$ of M at 0 is defined as the set of all (germs at 0) of holomorphic diffeomorphisms $H\colon (\mathbb{C}^N,0)\to(\mathbb{C}^N,0)$ such that $H(M)\subset M$, where the group structure is given by composition. The group $\mathrm{Hol}\,(M,0)$ is equipped with its natural inductive limit topology. In particular, a sequence $\{H^j\}$ in $\mathrm{Hol}\,(M,0)$ converges to H if there exists a compact neighborhood $\overline{U}$ of 0 such that each H^j has a representative holomorphic in an open neighborhood of $\overline{U}$ and H^j converges uniformly to H on $\overline{U}$.

One may ask whether $\mathrm{Hol}\,(M,0)$ may be given the structure of a finite dimensional Lie group compatible with its topology. For any integer $\ell\geq 0$, let $G^\ell(\mathbb{C}^N)$ denote the set of all ℓ-jets of invertible holomorphic mappings, i.e.

$$G^\ell(\mathbb{C}^N):=\{j_0^\ell H:\ H\colon(\mathbb{C}^N,0)\to(\mathbb{C}^N,0),\ H \text{ biholomorphic}\}. \tag{6.1}$$

The set $G^\ell(\mathbb{C}^N)$ has a finite dimensional Lie group structure with the multiplication defined by $(j_0^\ell H^1)\cdot(j_0^\ell H^2):=j_0^\ell(H^1\circ H^2)$. It is easy to check that this multiplication is independent of the choice of representatives H^1 and H^2. If M is holomorphically nondegenerate and of finite type at 0, then by taking $M=M'$ and $H^0=\mathrm{Id}$ in Theorem 5.2, it follows that there exists an integer $K>0$ such that the mapping $\mathrm{Hol}\,(M,0)\ni H\mapsto j_0^K H\in G^K(C^N)$ is continuous and injective. Hence $\mathrm{Hol}\,(M,0)$ may be identified with a subgroup of $G^K(\mathbb{C}^N)$ in this case. However, to show that $\mathrm{Hol}\,(M,0)$ is a Lie group one must show that its image is closed in $G^K(\mathbb{C}^N)$. This was proved for any finitely nondegenerate hypersurface M by the author with Baouendi and Ebenfelt [3] and later by Zaitsev [50] for any finitely nondegenerate generic submanifold of finite type in higher codimension. The following sharper result was given in [5].

Theorem 6.1 ([5]). *Let $M\subset\mathbb{C}^N$ be a real-analytic generic submanifold, which is ℓ_0-nondegenerate and of finite type at 0. Then the mapping*

$$\mathrm{Hol}\,(M,0)\ni H\mapsto j_0^{\ell_0(d+1)}H\in G^{\ell_0(d+1)}(\mathbb{C}^N),$$

taking a germ of a local biholomorphism at 0 to its $\ell_0(d+1)$-jet, gives a diffeomorphism of $\mathrm{Hol}\,(M,0)$ onto a real-algebraic Lie subgroup of $G^{\ell_0(d+1)}(\mathbb{C}^N)$.

It should be noted here that Theorem 6.1 for the case of a Levi-nondegenerate hypersurface (i.e. $\ell_0=d=1$) follows from the work of Chern–Moser [20] and Burns–Shnider [18].

Recent work of Kim–Zaitsev [38] gives a construction of a smooth hypersurface $M\subset\mathbb{C}^N$, finitely nondegenerate at 0, for which $\mathrm{Hol}\,(M,0)$ is not a Lie group, although it is contained as a subgroup of a finite dimensional Lie group. In attempting to generalize Theorem 6.1 in another direction, one may ask whether the condition that M be finitely nondegenerate can be weakened (see the equivalences (i)–(vii) in Section 3). As observed above, in the case of a holomorphically nondegenerate generic submanifold of finite type there is an integer K for which the mapping

$\mathrm{Hol}\,(M,0) \ni H \mapsto j_0^K H \in G^K(\mathbb{C})$ is an injection, but it is not known if its image is closed. However, very recently Lamel and Mir [42] gave a positive answer in the slightly more restrictive case of a real-analytic generic submanifold that is essentially finite and of finite type at 0.

Although global questions are outside the announced scope of this paper, I would like to mention here a recent result joint with Baouendi, Winkelmann, and Zaitsev [16] in which we prove that the global automorphism group of a CR manifold that is finitely nondegenerate and of finite type at every point has the structure of a finite dimensional Lie group. This work uses the above mentioned results of Kim–Zaitsev [39].

7. Algebraicity of mappings

Recall that a real submanifold is *real-algebraic* if it is contained in a real-algebraic subset of the same dimension. A (germ at 0 of a) holomorphic mapping $H\colon (\mathbb{C}^N,0) \to (\mathbb{C}^N,0)$ is *complex-algebraic* if its graph is a (germ at 0 of a) complex algebraic variety of $\mathbb{C}^N \times \mathbb{C}^N$. One may ask under what conditions a mapping that sends one real-algebraic generic submanifold into another is necessarily complex-algebraic. Very early results on this question are contained in the work of Poincaré [46], who proved that a local biholomorphism between two pieces of spheres in $\mathbb{C}^2$ must be a rational mapping. In 1977 Webster [48] solved the above problem for the case when M and M' are algebraic hypersurfaces that are Levi-nondegenerate, proving that any such local biholomorphism is necessarily complex-algebraic. The following result for higher codimensional submanifolds was obtained in joint work of the author with Baouendi and Ebenfelt [2].

Theorem 7.1 ([2]). *Let $M, M' \subset \mathbb{C}^N$ be two real-algebraic generic submanifolds of the same dimension through 0. Assume that M is connected and of finite type and holomorphically nondegenerate at some point. Then if $H\colon (\mathbb{C}^N,0) \to (\mathbb{C}^N,0)$ is a germ of a holomorphic mapping with $H(M) \subset M'$ satisfying $\mathrm{Jac}\, H \not\equiv 0$ then H is complex-algebraic.*

If the condition that $\mathrm{Jac}\, H \not\equiv 0$ is not assumed, or if M and M' are generic submanifolds in complex spaces of different dimensions, a stronger assumption must be imposed on the target space. For the case of strongly pseudoconvex hypersurfaces of different dimensions, see the work of Huang [37], see also [10]. One of the most general results in this direction is due to Zaitsev:

Theorem 7.2 (Zaitsev [51]). *Let $M \subset \mathbb{C}^N$ and $M' \subset \mathbb{C}^{N'}$ be real-algebraic submanifolds through 0. Then all (germs at 0 of) local holomorphic maps $H\colon (\mathbb{C}^N,0) \to (\mathbb{C}^{N'},0)$ with $H(M) \subset M'$ are complex-algebraic if and only if the following are satisfied:*

(i) *M is generic and of finite type on a dense subset;*

(ii) *M' contains no analytic discs.*

8. Concluding remarks

Although I have attempted to survey a number of recent results in CR geometry that make use of the Segre mappings (as defined in Section 2), I have omitted a number of other interesting questions to which this method has also been applied. In particular, I would like to mention one such area of current research, namely the study of the convergence of formal mappings sending one generic submanifold into another. Here "formal" means that the mappings consist of formal power series, rather than convergent ones. In addition to the method of Segre mappings, the celebrated Artin Approximation Theorem [1] has been an important tool in this area of research. For some recent results in this direction I refer the reader to [7], [44], [45], [41], [13], [17], [43]. With these references, as well as with all the other references given here, I apologize in advance for any omissions.

References

[1] Artin, M., On the solutions of analytic equations. *Invent. Math.* **5** (1968), 277–291.

[2] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., Algebraicity of holomorphic mappings between real algebraic sets in $\mathbf{C}^n$. *Acta Math.* **177** (2) (1996), 225–273.

[3] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., Parametrization of local biholomorphisms of real analytic hypersurfaces. *Asian J. Math.* **1** (1) (1997), 1–16.

[4] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., CR automorphisms of real analytic manifolds in complex space. *Comm. Anal. Geom.* **6** (2) (1998), 291–315.

[5] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., Rational dependence of smooth and analytic CR mappings on their jets. *Math. Ann.* **315** (2) (1999), 205–249.

[6] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., *Real submanifolds in complex space and their mappings*. Princeton Math. Ser. 47, Princeton University Press, Princeton, NJ, 1999.

[7] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., Convergence and finite determination of formal CR mappings. *J. Amer. Math. Soc.* **13** (4) (2000), 697–723.

[8] Baouendi, M. S., Ebenfelt, P., and Rothschild, L. P., Dynamics of the Segre varieties of a real submanifold in complex space. *J. Algebraic Geom.* **12** (1) (2003), 81–106.

[9] Baouendi, M. S., and Huang, X., Super-rigidity for holomorphic mappings between hyperquadrics with positive signature. *J. Differential Geom.* **69** (2) (2005), 379–398.

[10] Baouendi, M. S., Huang, X., and Rothschild, P. L., Regularity of CR mappings between algebraic hypersurfaces. *Invent. Math.* **125** (1) (1996), 13–36.

[11] Baouendi, M. S., Huang, X. J., and Rothschild, L. P., Nonvanishing of the differential of holomorphic mappings at boundary points. *Math. Res. Lett.* **2** (6) (1995), 737–750.

[12] Baouendi, M. S., Jacobowitz, H., and Treves, F., On the analyticity of CR mappings. *Ann. of Math.* (2) **122** (2) (1985), 365–400.

[13] Baouendi, M. S., Mir, N., and Rothschild, L. P., Reflection ideals and mappings between generic submanifolds in complex space. *J. Geom. Anal.* **12** (4) (2002), 543–580.

[14] Baouendi, M. S., and Rothschild, L. P., Geometric properties of mappings between hypersurfaces in complex space. *J. Differential Geom.* **31** (2) (1990), 473–499.

[15] Baouendi, M. S., and Rothschild, L. P., A generalized complex Hopf lemma and its applications to CR mappings. *Invent. Math.* **111** (2) (1993), 331–348.

[16] Baouendi, M. S., Rothschild, L. P., Winkelmann, J., and Zaitsev, D., Lie group structures on groups of diffeomorphisms and applications to CR manifolds. *Ann. Inst. Fourier* (*Grenoble*) **54** (5) (2004), 1279–1303.

[17] Baouendi, M. S., Rothschild, L. P., and Zaitsev, D., Equivalences of real submanifolds in complex space. *J. Differential Geom.* **59** (2) (2001), 301–351.

[18] Burns, Jr., D., and Shnider, S., Real hypersurfaces in complex manifolds. In *Several complex variables* (Williams Coll., Williamstown, Mass., 1975), Part 2, Proc. Sympos. Pure Math. XXX, Amer. Math. Soc., Providence, R.I., 1977, 141–168.

[19] Chern, S.-S., and Ji, S., On the Riemann mapping theorem. *Ann. of Math.* (2) **144** (2) (1996), 421–439.

[20] Chern, S. S., and Moser, J. K., Real hypersurfaces in complex manifolds. *Acta Math.* **133** (1974), 219–271.

[21] Christ, M., Nagel, A., Stein, E. M., and Wainger, S., Singular and maximal Radon transforms: analysis and geometry. *Ann. of Math.* (2) **150** (2) (1999), 489–577.

[22] D'Angelo, J. P., Real hypersurfaces, orders of contact, and applications. *Ann. of Math.* (2) **115** (3) (1982), 615–637.

[23] D'Angelo, J. P., The geometry of proper holomorphic maps between balls. In *The Madison Symposium on Complex Analysis* (Madison, WI, 1991), Contemp. Math. 137, Amer. Math. Soc., Providence, RI, 1992, 191–215.

[24] Diederich, K., and Pinchuk, S., Reflection principle in higher dimensions. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 703–712.

[25] Diederich, K., and Webster, S. M., A reflection principle for degenerate real hypersurfaces. *Duke Math. J.* **47** (4) (1980), 835–843.

[26] Ebenfelt, P., Finite jet determination of holomorphic mappings at the boundary. *Asian J. Math.* **5** (4) (2001), 637–662.

[27] Ebenfelt, P., Huang, X., and Zaitsev, D., Rigidity of CR-immersions into spheres. *Comm. Anal. Geom.* **12** (3) (2004), 631–670.

[28] Ebenfelt, P., and Lamel, B., Finite jet determination of CR embeddings. *J. Geom. Anal.* **14** (2) (2004), 241–265.

[29] Ebenfelt, P., Lamel, B., and Zaitsev, D., Finite jet determination of local analytic CR automorphisms and their parametrization by 2-jets in the finite type case. *Geom. Funct. Anal.* **13** (3) (2003), 546–573.

[30] Ebenfelt, P., and Rothschild, L. P., Transversality of CR mappings. *Amer. J. Math.*, to appear.

[31] Fornaess, J. E., Embedding strictly pseudoconvex domains in convex domains. *Amer. J. Math.* **98** (2) (1976), 529–569.

[32] Fornaess, J. E., Biholomorphic mappings between weakly pseudoconvex domains. *Pacific J. Math.* **74** (1) (1978), 63–65.

[33] Forstnerič, F., Proper holomorphic mappings: a survey. In *Several complex variables* (Stockholm, 1987/1988), Math. Notes 38, Princeton University Press, Princeton, NJ, 1993, 297–363.

[34] Han, C.-K., Complete differential system for the mappings of CR manifolds of nondegenerate Levi forms. *Math. Ann.* **309** (3) (1997), 401–409.

[35] Huang, X., On a linearity problem for proper holomorphic maps between balls in complex spaces of different dimensions. *J. Differential Geom.* **51** (1) (1999), 13–33.

[36] Huang, X., and Ji, S., Mapping $\mathbf{B}^n$ into $\mathbf{B}^{2n-1}$. *Invent. Math.* **145** (2) (2001), 219–250.

[37] Huang, X. J., On the mapping problem for algebraic real hypersurfaces in the complex spaces of different dimensions. *Ann. Inst. Fourier* (*Grenoble*) **44** (2) (1994), 433–463.

[38] Kim, S.-Y., and Zaitsev, D., Remarks on the rigidity of CR-manifolds. Preprint.

[39] Kim, S.-Y., and Zaitsev, D., Equivalence and embedding problems for CR-structures of any codimension. *Topology* **44** (3) (2005), 557–584.

[40] Kohn, J. J., Boundary behavior of δ on weakly pseudo-convex manifolds of dimension two. *J. Differential Geometry* **6** (1972), 523–542.

[41] Lamel, B., Holomorphic maps of real submanifolds in complex spaces of different dimensions. *Pacific J. Math.* **201** (2) (2001), 357–387.

[42] Lamel, B., and Mir, N., Parametrization of local CR automorphisms by finite jets and applications. *J. Amer. Math. Soc.*, to appear.

[43] Meylan, F., Mir, N., and Zaitsev, D., Approximation and convergence of formal CR-mappings. *Internat. Math. Res. Notices* **2003** (4) (2003), 211–242.

[44] Mir, N., Formal biholomorphic maps of real analytic hypersurfaces. *Math. Res. Lett.* **7** (2–3) (2000), 343–359.

[45] Mir, N., Convergence of formal embeddings between real-analytic hypersurfaces in codimension one. *J. Differential Geom.* **62** (1) (2002), 163–173.

[46] Poincaré, H., Les fonctions analytiques de deux variables et la représentation conforme. *Rend. Circ. Mat. Palermo* **23** (1907), 185–220.

[47] Stanton, N. K., Infinitesimal CR automorphisms of rigid hypersurfaces. *Amer. J. Math.* **117** (1) (1995), 141–167.

[48] Webster, S. M., On the mapping problem for algebraic real hypersurfaces. *Invent. Math.* **43** (1) (1977), 53–68.

[49] Webster, S. M., On the reflection principle in several complex variables. *Proc. Amer. Math. Soc.* **71** (1) (1978), 26–28.

[50] Zaitsev, D., Germs of local automorphisms of real-analytic CR structures and analytic dependence on k-jets. *Math. Res. Lett.* **4** (6) (1997), 823–842.

[51] Zaitsev, D., Algebraicity of local holomorphisms between real-algebraic submanifolds of complex spaces. *Acta Math.* **183** (2) (1999), 273–305.

[52] Zaitsev, D., Unique determination of local CR-maps by their jets: a survey. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei* (9) *Mat. Appl.* **13** (3–4) (2002), 295–305.

Department of Mathematics, University of California at San Diego, La Jolla,
CA 92093-0112, U.S.A.

E-mail: lrothschild@ucsd.edu

Towards conformal invariance of 2D lattice models

Stanislav Smirnov

Abstract. Many 2D lattice models of physical phenomena are conjectured to have conformally invariant scaling limits: percolation, Ising model, self-avoiding polymers, etc. This has led to numerous exact (but non-rigorous) predictions of their scaling exponents and dimensions. We will discuss how to prove the conformal invariance conjectures, especially in relation to Schramm–Loewner evolution.

Mathematics Subject Classification (2000). Primary 82B20; Secondary 60K35, 82B43, 30C35, 81T40.

Keywords. Statistical physics, conformal invariance, universality, Ising model, percolation, SLE.

1. Introduction

For several 2D lattice models physicists were able to make a number of spectacular predictions (non-rigorous, but very convincing) about exact values of various scaling exponents and dimensions. Many methods were employed (Coulomb gas, Conformal Field Theory, Quantum Gravity) with one underlying idea: that the model at criticality has a continuum scaling limit (as mesh of the lattice goes to zero) and the latter is conformally invariant. Moreover, it is expected that there is only a one-parameter family of possible conformally invariant scaling limits, so universality follows: if the same model on different lattices (and sometimes at different temperatures) has a conformally invariant scaling limit, it is necessarily the same. Indeed, the two limits belong to the same one-parameter family, and usually it directly follows that the corresponding parameter values coincide.

Recently mathematicians were able to offer different, perhaps better, and certainly more rigorous understanding of those predictions, in many cases providing proofs. The point which is perhaps still less understood both from mathematics and physics points of view is why there exists a universal conformally invariant scaling limit. However such behavior is supposed to be typical in 2D models at criticality: Ising, percolation, self-avoiding polymers; with universal conformally invariant curves arising as scaling limits of the interfaces.

Until recently this was established only for the scaling limit of the 2D random walk, the 2D Brownian motion. This case is easier and somehow exceptional because of the Markov property. Indeed, Brownian motion was originally constructed

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

by Wiener [43], and its conformal invariance (which holds in dimension 2 only) was shown by Paul Lévy [24] without appealing to random walk. Note also that unlike interfaces (which are often simple, or at most "touch" themselves), Brownian trajectory has many "transversal" self-intersections.

For other lattice models even a rigorous formulation of conformal invariance conjecture seemed elusive. Considering percolation (a model where vertices of a graph are declared open independently with equal probability p – see the discussion below) at criticality as an example, Robert Langlands, Philippe Pouliot and Yvan Saint-Aubin in [20] studied numerically crossing probabilities (of events that there is an open crossing of a given rectangular shape). Based on experiments they concluded that crossing probabilities should have a universal (independent of lattice) scaling limit, which is conformally invariant (a conjecture they attributed to Michael Aizenman). Thus the limit of crossing probability for a rectangular domain should depend on its conformal modulus only. Moreover an exact formula (5) using hypergeometric function was proposed by John Cardy in [9] based on Conformal Field Theory arguments. Later Lennart Carleson found that the formula has a particularly nice form for equilateral triangles, see [38]. These developments got many researchers interested in the subject and stimulated much of the subsequent progress.

Rick Kenyon [16], [17] established conformal invariance of many observables related to dimer models (domino tilings), in particular to uniform spanning tree and loop erased random walk, but stopped short of constructing the limiting curves.

In [32], Oded Schramm suggested to study the scaling limit of a single interface and classified all possible curves which can occur as conformally invariant scaling limits. Those turned out to be a universal one-parameter family of SLE(κ) curves, which are now called *Schramm–Loewner evolutions*. The word "evolution" is used since the curves are constructed dynamically, by running classical Loewner evolution with Brownian motion as a driving term. We will discuss one possible setup, chordal SLE(κ) with parameter $\kappa \in [0, \infty)$, which provides for each simply-connected domain Ω and boundary points a, b a measure μ on curves from a to b inside Ω. The measures $\mu(\Omega, a, b)$ are conformally invariant, in particular they are all images of one measure on a reference domain, say a half-plane $\mathbb{C}_+$. An exact definition appears below.

In [37], [38] the conformal invariance was established for critical percolation on triangular lattice. Conformally invariant limit of the interface was identified with SLE(6), though its construction does not use SLE machinery. See also Federico Camia and Charles Newman's paper [8] for the details on subsequent construction of the full scaling limit.

In [23] Greg Lawler, Oded Schramm and Wendelin Werner have shown that a perimeter curve of the uniform spanning tree converges to SLE(8) (and the related loop erased random walk – to SLE(2)) on a general class of lattices. Unlike the proof for percolation, theirs utilizes SLE in a substantial way. In [35], Oded Schramm and Scott Sheffield introduced a new model, Harmonic Explorer, where properties needed for convergence to SLE(4) are built in.

Despite the results for percolation and uniform spanning tree, the problem remained open for all other classical (spin and random cluster) 2D models, including percolation on other lattices. This was surprising given the abundance of the physics literature on conformal invariance. Perhaps most surprising was that the problem of a conformally invariant scaling limit remained open for the Ising model, since for the latter there are many exact and often rigorous results – see the books [27], [6].

Recently we were able to work out the Ising case [39]:

Theorem 1. *As lattice step goes to zero, interfaces in Ising and Ising random cluster models on the square lattice at critical temperature converge to* SLE(3) *and* SLE(16/3) *correspondingly.*

Computer simulations of these interfaces (Figures 2, 4) as well as the definition of the Ising models can be found below. Similarly to mentioned experiments for percolation, Robert Langlands, Marc André Lewis and Yvan Saint-Aubin conducted in [21] numerical studies of crossing probabilities for the Ising model at critical temperature. A modification of the theorem above relating interfaces to SLE's (with drifts) in domains with five marked boundary points allows a rigorous setup for establishing their conjectures.

The proof is based on showing that a certain Fermionic lattice observable (or rather two similar ones for spin and random cluster models) is discrete analytic and solves a particular covariant Riemann Boundary Value Problem. Hence its limit is conformally covariant and can be calculated exactly. The statement is interesting in its own right, and can be used to study spin correlations. The observable studied has more manifest physics meaning than one in our percolation paper [38].

The methods lead to some progress in fairly general families of random cluster and $O(n)$ models, and not just on square lattices. In particular, besides Ising cases, they seem to suggest new proofs for all other known cases (i.e. site percolation on triangular lattice and uniform spanning tree).

In this note we will discuss this proof and general approach to scaling limits and conformal invariance of interfaces in the SLE context. We will also state some of the open questions and speculate on how one should approach other models.

We omit many aspects of this rich subject. We do not discuss the general mathematical theory of SLE curves or their connections to physics, for which interested reader can consult the expository works [5], [10], [15], [41] and the book [22]. We do not mention the question of how to deduce the values of scaling exponents for lattice models with SLE help once convergence is known. It was explored in some detail only for percolation [40], where convergence is known and the required (difficult) estimates were already in place thanks to Harry Kesten [19]. We also restrict ourselves to one interface, whereas one can study the collection of all loops (cf. exposition [42]), and many of our considerations transfer to the loop soup observables. Finally, there are many other open questions related to conformal invariance, some of which are discussed in Oded Schramm's paper [34] in these proceedings.

2. Lattice models

We focus on two families of lattice models which have nice "loop representations". Those families include or are closely related to most of the "important" models, including percolation, Ising, Potts, spherical (or $O(n)$), Fortuin–Kasteleyn (or random cluster), self-avoiding random walk, and uniform spanning tree models. For their interrelations and for the discussion of many other relevant models one can consult the books [6], [12], [26], [27]. We also omit many references which can be found there.

There are various ways to understand the existence of the scaling limit and its conformal invariance. One can ask for the full picture, which can be represented as a loop collection (representing all cluster interfaces), random height function (changing by ± 1 whenever we cross a loop), or some other object. It however seems desirable to start with a simpler problem.

One can start with observables (like correlation functions, crossing probabilities), for which it is easier to make sense of the limit: there should exist a limit of a number sequence which is a conformal invariant. Though a priori it might seem to be a weaker goal than constructing a full scaling limit, there are indications that to obtain the full result it might be sufficient to analyze just one observable.

We will discuss an intermediate goal to analyze the law of just one interface, explain why working out just one observable would be sufficient, and give details on how to find an observable with a conformally invariant limit. To single out one interface, we consider a model on a simply connected domain with Dobrushin boundary conditions (which besides many loop interfaces enforce existence of an interface joining two boundary points a and b). We omit the discussion of the full scaling limit, as well as models on Riemann surfaces and with different boundary conditions.

2.1. Percolation. Perhaps the simplest model (to state) is Bernoulli percolation on the triangular lattice. Vertices are declared open or closed (grey or white in Figure 1) independently with probabilities p and $(1-p)$ correspondingly. The critical value is $p = p_c = 1/2$ – see [18], [11], in which case all colorings are equally probable.

Then each configuration can be represented by a collection of interfaces – loops which go along the edges of the dual hexagonal lattice and separate open and closed vertices.

We want to distinguish one particular interface, and to this effect we introduce Dobrushin boundary conditions: we take two boundary points a and b in a simply connected Ω (or rather its lattice approximation), asking the counterclockwise arc ab to be grey and the counterclockwise arc ba to be white. This enforces existence of a single non-loop interface which runs from a to b. The "loop gas" formulation of our model is that we consider all collections of disjoint loops plus a curve from a to b on hexagonal lattice with equal probability.

For each value of the lattice step $\varepsilon > 0$ we approximate a given domain Ω by a lattice domain, which leads to a random interface, that is a probability measure μ_ε

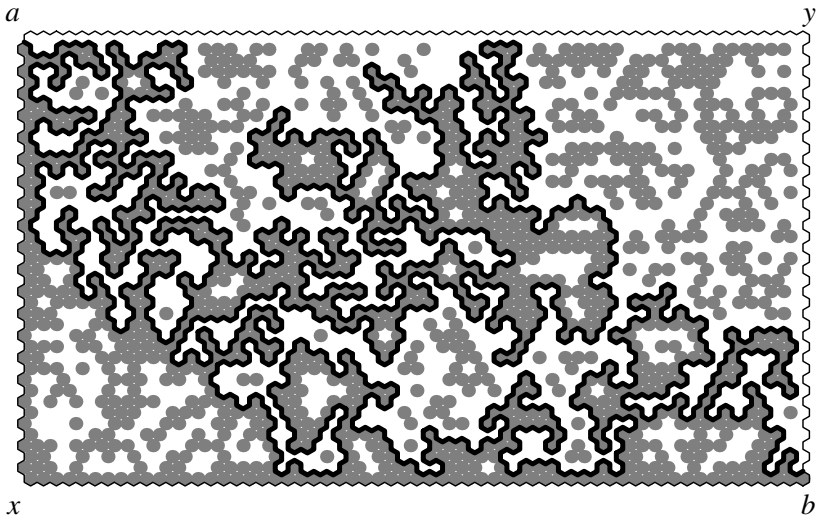


Figure 1. Critical site percolation on triangular lattice superimposed over a rectangle. Every site is grey or white independently with equal probability 1/2. Dobrushin boundary conditions (grey on lower and left sides, white on upper and right sides) produce an interface from the upper left corner a to the lower right corner b. The law of the interface converges to SLE(6) when lattice step goes to zero, while the rectangle is fixed.

on curves (broken lines) running from a to b. The question is whether there is a limit measure $\mu = \mu(\Omega, a, b)$ on curves and whether it is conformally invariant. To make sense of the limit we consider the curves with uniform topology generated by parameterizations (with distance between γ_1 and γ_2 being $\inf \|f_1 - f_2\|_\infty$ where the infimum is taken over all parameterizations f_1, f_2 of γ_1, γ_2), and ask for weak-$*$ convergence of the measures μ_ε.

2.2. $O(n)$ and loop models. Percolation turns out to be a particular case of the *loop gas* model which is closely related (via high-temperature expansion) to $O(n)$ (spherical) model. We consider configurations of non-intersecting simple loops and a curve running from a to b on *hexagonal lattice* inside domain Ω as for percolation in Figure 1. But instead of asking all configurations to be equally likely, we introduce two parameters: loop-weight $n \geq 0$ and edge-weight $x > 0$, and ask that probability of a configuration is proportional to

$$n^{\#\text{ loops}}\, x^{\text{length of loops}}.$$

The vertices not visited by loops are called monomers. Instead of weighting edges by x one can equivalently weight monomers by $1/x$.

We are interested in the range $n \in [0, 2]$ (after certain modifications $n \in [-2, 2]$ would work), where conformal invariance is expected (other values of n have different

behavior). It turns out that there is a critical value $x_c(n)$, such that the model exhibits one critical behavior at $x_c(n)$ and another on the interval $(x_c(n), +\infty)$, corresponding to "dilute" and "dense" phases (when in the limit the loops are simple and non-simple correspondingly).

Bernard Nienhuis [28], [29] proposed the following conjecture, supported by physics arguments:

Conjecture 2. The critical value is given by

$$x_c(n) = \frac{1}{\sqrt{2+\sqrt{2-n}}}.$$

Note that though for all $x \in (x_c(n), \infty)$ the critical behavior (and the scaling limit) are conjecturally the same, the related value $\tilde{x}_c(n) = 1/\sqrt{2-\sqrt{2-n}}$ turns out to be distinguished in some ways.

The criticality was rigorously established for $n = 1$ only, but we still may discuss the scaling limits at those values of x. It is widely believed that at the critical values the model has a conformally invariant scaling limit. Moreover, the corresponding criticalities under renormalization are supposed to be unstable and stable correspondingly, so for $x = x_c$ there should be one conformally invariant scaling limit, whereas for the interval $x \in (x_c, \infty)$ another, corresponding to $\tilde{x}_c$. The scaling limit for low temperatures $x \in (0, x_c)$, a straight segment, is not conformally invariant.

Plugging in $n = 1$ we obtain weight

$$x^{\text{length of loops}}.$$

Assigning the spins ± 1 (represented by grey and white colors in Figure 1) to sites of triangular lattice, we rewrite the weight as

$$x^{\#\text{ pairs of neighbors of opposite spins}}, \tag{1}$$

obtaining the Ising model (where the usual parameterization is $\exp(-2\beta) = x$). The critical value is known to be $\beta_c = \log 3/4$, so one gets the Ising model at critical temperature for $n = 1$, $x = 1/\sqrt{3}$. A computer simulation of the Ising model on the square lattice at critical temperature, when the probability of configuration is proportional to (1), is shown in Figure 2.

For $n = 1$, $x = 1$ we obtain critical site percolation on triangular lattice. Taking $n = 0$ (which amounts to considering configurations with no loops, just a curve running from a to b), one obtains for $x_c = 1/\sqrt{2+\sqrt{2}}$ a version of the self-avoiding random walk.

The following conjecture (see e.g. [15]) is a direct consequence of physics predictions and SLE calculations:

Conjecture 3. For $n \in [0, 2]$ and $x = x_c(n)$, as lattice step goes to zero, the law of the interface converges to Schramm–Loewner evolution with

$$\kappa = 4\pi/(2\pi - \arccos(-n/2)).$$

For $n \in [0, 2]$ and $x \in (x_c, \infty)$ (in particular for $x = \tilde{x}_c$), as lattice step goes to zero, the law of the interface converges to Schramm–Loewner evolution with

$$\kappa = 4\pi / \arccos(-n/2).$$

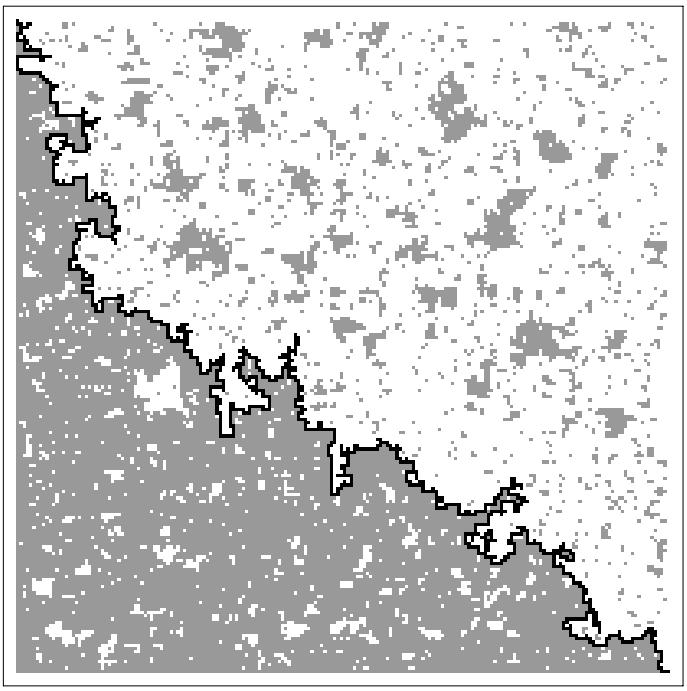


Figure 2. The Ising model at critical temperature on a square. White and grey sites represent ± 1 spins. Dobrushin boundary conditions (grey on lower and left sides, white on upper and right sides) produce, besides loop interfaces, an interface from the upper left to the lower right corner, pictured in black. When lattice step goes to zero, the law of the interface converges to SLE(3), which is a conformally invariant random curve, almost surely simple and of Hausdorff dimension 11/8.

Note that to address this question one does not need to prove that the Nienhuis temperature is indeed critical (Conjecture 2).

We discussed loops on the hexagonal lattice, since it is a trivalent graph and so at most one interface can pass through a vertex. One can engage in similar considerations on the square lattice with special regard to a possibility of two interfaces passing through the same vertex, in which case they can be split into loops in two different ways (with different configuration weights). In the case of Ising ($n = 1$) this poses less of a problem, since number of loops is not important. For $n = 1$ and $x = 1$ we get percolation model with $p = 1/2$, but for a general lattice this p need not be critical, so e.g. critical site percolation on the square lattice does not fit directly into this framework.

2.3. Fortuin–Kasteleyn random cluster models. Another interesting class is Fortuin–Kasteleyn models, which are random cluster representations of q-state Potts model. The random cluster measure on a graph (a piece of the *square lattice* in our case) is a probability measure on edge configurations (each edge is declared either open or closed), such that the probability of a configuration is proportional to

$$p^{\#\text{ open edges}}\,(1-p)^{\#\text{ closed edges}}\,q^{\#\text{ clusters}},$$

where clusters are maximal subgraphs connected by open edges. The two parameters are edge-weight $p \in [0,1]$ and cluster-weight $q \in (0,\infty)$, with $q \in [0,4]$ being interesting in our framework (similarly to the previous model, $q > 4$ exhibits different behavior). For a square lattice (or in general any planar graph) to every configuration one can prescribe a cluster configuration on the dual graph, such that every open edge is intersected by a dual closed edge and vice versa. See Figure 3 for a picture of two dual configurations with respective open edges. It turns out that the probability of a dual configuration becomes proportional to

$$p_*^{\#\text{ dual open edges}}\,(1-p_*)^{\#\text{ dual closed edges}}\,q^{\#\text{ dual clusters}},$$

with the dual to p value $p_* = p_*(p)$ satisfying $p_*/(1-p_*) = q(1-p)/p$. For $p = p_{\mathrm{sd}} := \sqrt{q}/(\sqrt{q}+1)$ the dual value coincides with the original one: one gets $p_{\mathrm{sd}} = (p_{\mathrm{sd}})_*$ and so the model is self-dual. It is conjectured that this is also the critical value of p, which was only proved for $q = 1$ (percolation), $q = 2$ (Ising) and $q > 25.72$.

Again we introduce Dobrushin boundary conditions: wired on the counterclockwise arc ab (meaning that all edges along the arc are open) and dual-wired on the counterclockwise arc ba (meaning that all dual edges along the arc are open, or equivalently all primal edges orthogonal to the arc are closed) – see Figure 3. Then there is a unique interface running from a to b, which separates cluster containing the arc ab from the dual cluster containing the arc ba.

We will work with the loop representation, which is similar to that in 2.2. The cluster configurations can be represented as Hamiltonian (i.e. including all edges) non-intersecting (more precisely, there are no "transversal" intersections) loop configurations on the medial lattice. The latter is a square lattice which has edge centers of the original lattice as vertices. The loops represent interfaces between cluster and dual clusters and turn by $\pm\frac{\pi}{2}$ at every vertex – see Figure 3. It is well-known that probability of a configuration is proportional to

$$\left(\frac{p}{1-p}\frac{1}{\sqrt{q}}\right)^{\#\text{ open edges}} \cdot \left(\sqrt{q}\right)^{\#\text{ loops}},$$

which for the self-dual value $p = p_{\mathrm{sd}}$ simplifies to

$$\left(\sqrt{q}\right)^{\#\text{ loops}}. \tag{2}$$

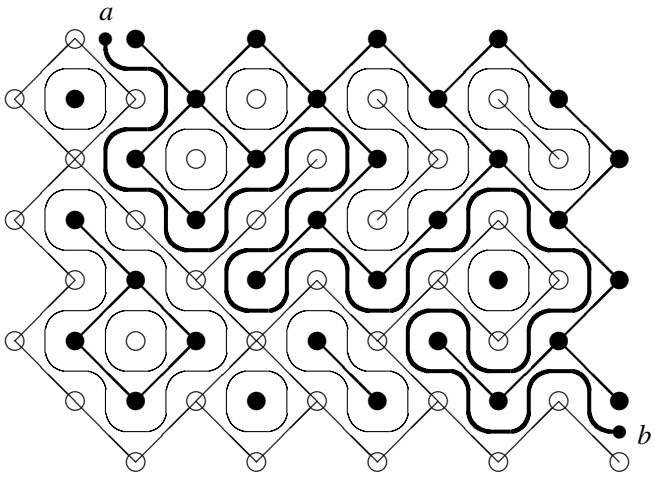


Figure 3. Loop representation of the random cluster model. The sites of the original lattice are colored in black, while the sites of the dual lattice are colored in white. Clusters, dual clusters and loops separating them are pictured. Under Dobrushin boundary conditions besides a number of loops there is an interface running from a to b, which is drawn in bold. Weight of the configuration is proportional to $(\sqrt{q})^{\#\text{ loops}}$.

Dobrushin boundary conditions amount to introducing two vertices with odd number of edges: a source a and a sink b, which enforces a curve running form a to b (besides loops) – see Figure 3 for a typical configuration.

Conjecture 4. For all $q \in [0, 4]$, as the lattice step goes to zero, the law of the interface converges to Schramm–Loewner evolution with $\kappa = 4\pi/\arccos(-\sqrt{q}/2)$.

The conjecture was proved by Greg Lawler, Oded Schramm and Wendelin Werner [23] for the case of $q = 0$, when they showed that the perimeter curve of the uniform spanning tree converges to SLE(8). Note that with Dobrushin boundary conditions loop representation still makes sense for $q = 0$. In fact, the formula (2) means that we restrict ourselves to configurations with no loops, just a curve running from a to b (which then necessarily passes through all the edges), and all configurations are equally probable.

Below we will outline our proof [39] that for the Ising parameter $q = 2$ the interface converges to SLE(16/3), see Figure 4. It almost directly translates into a proof that the interface of the spin cluster for the Ising model on the square lattice at the critical temperature (which can be rewritten as the loop model in 2.2 for $n = 1$, only on the square lattice) converges to SLE(3), as shown in Figure 2. It seems likely that it will work in the $n = 1$ case for the loop model on hexagonal lattice described above, providing convergence to SLE(3) for $x = x_c$ and (a new proof of) convergence to SLE(6) for $x = \tilde{x}_c$ (and possibly for all $x > x_c$).

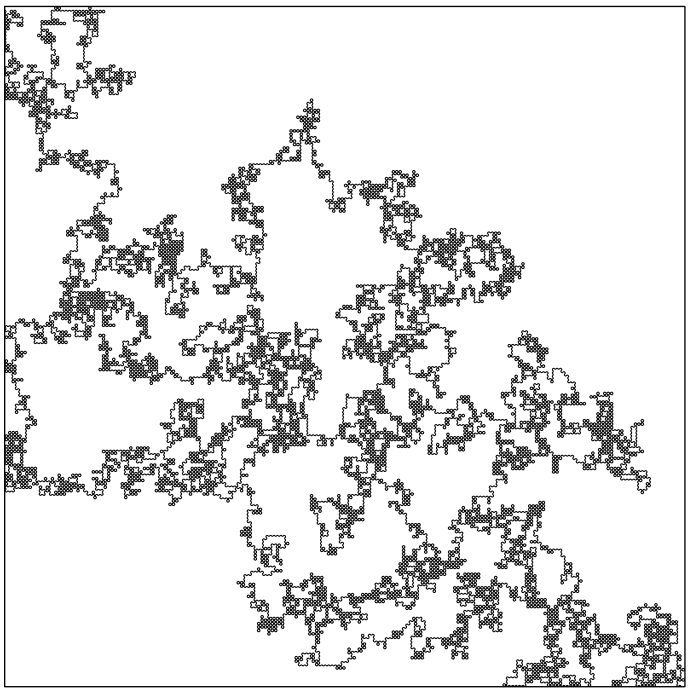


Figure 4. Interface in the random cluster Ising model at critical temperature with Dobrushin boundary conditions (loops not pictured). The law converges to SLE(16/3) when mesh goes to zero, so in the limit it has Hausdorff dimension $5/3$ and touches itself almost surely. The random cluster is obtained by deleting some bonds from the spin cluster, so the interfaces are naturally different. Indeed, they converge to different SLE's and have different dimensions. However they are related: conjecturally, the outer boundary of the (non-simple) pictured curve and the (simple) spin interface in Figure 2 have the same limit after appropriate conditioning.

Summing it up, Conjecture 3 was proved earlier for $n = 1$, $x = \tilde{x}_c$, see [38], [37], whereas Conjecture 4 was established for $q = 0$, see [23]. We outline a technique, which seems to prove conformal invariance in two new cases, and provide new proofs for the only cases known before, making Conjecture 4 solved for $q = 0$ and $q = 2$, and Conjecture 3 for $n = 1$. The method also contributes to our understanding of universality phenomenon.

Much of the method works for general values of n and q. The most interesting values of the parameters (where it does not yet work all the way) are $n = 0$, related to the self-avoiding random walk, and $q = 1$, equivalent to the critical bond percolation on the square lattice (in the latter case some progress was achieved by Vincent Beffara by a different method). Hopefully the lemma (essentially the discrete analyticity statement – see below) required to transfer our proof to other models will be worked out someday, leading to full resolution of these conjectures.

3. Schramm–Loewner evolution

3.1. Loewner evolution. Loewner evolution is a differential equation for a Riemann uniformization map for a domain with a growing slit. It was introduced by Charles Loewner in [25] in his work on Bieberbach's conjecture.

In the original work, Loewner considered slits growing towards interior point. Though such *radial* evolution (along with other possible setups) is also important in the context of lattice models and fits equally well into our framework, we will restrict ourselves to the *chordal* case, when the slit is growing towards a point on the boundary.

In both cases we choose a particular Riemann map by fixing its value and derivative at the target point. *Chordal Loewner evolution* describes uniformization for the upper half-plane $\mathbb{C}_+$ with a slit growing from 0 to ∞ (one deals with a general domain Ω with boundary points a, b by mapping it to $\mathbb{C}_+$ so that $a \mapsto 0$, $b \mapsto \infty$).

Loewner only considered slits given by smooth simple curves, but more generally one allows any set which grows continuously in conformal metric when viewed from ∞. We will omit the precise definition of *allowed slits* (more extensive discussion in this context can be found in [22]), only noting that all simple curves are included. The random curves arising from lattice models (e.g. cluster perimeters or interfaces) are simple (or can be made simple by altering them on the local scale). Their scaling limits are not necessarily simple, but they have no "transversal" self-intersections. For such a curve to be an allowed slit it is sufficient if it touches itself to never venture into the created loop. This property would follow if e.g. a curve visits no point thrice.

Parameterizing the slit γ in some way by time t, we denote by $g_t(z)$ the conformal map sending $\mathbb{C}_+ \setminus \gamma_t$ (or rather its component at ∞) to $\mathbb{C}_+$ normalized so that at infinity $g_t(z) = z + \alpha(t)/z + \mathcal{O}(1/|z|^2)$, the so called *hydrodynamic normalization*. It turns out that $\alpha(t)$ is a continuous strictly increasing function (it is a sort of capacity-type parameter for γ_t), so one can change the time so that

$$g_t(z) = z + \frac{2t}{z} + \mathcal{O}\left(\frac{1}{|z|^2}\right). \tag{3}$$

Denote by $w(t)$ the image of the tip $\gamma(t)$. The family of maps g_t (also called a *Loewner chain*) is uniquely determined by the real-valued *"driving term"* $w(t)$. The general Loewner theorem can be roughly stated as follows:

Loewner's theorem. *There is a bijection between allowed slits and continuous real valued functions $w(t)$ given by the ordinary differential equation*

$$\partial_t g_t(z) = \frac{2}{g_t(z) - w(t)}, \quad g_0(z) = z. \tag{4}$$

The original Loewner equation is different since he worked with smooth radial slits and evolved them in another (but related) way.

3.2. Schramm–Loewner evolution. While a deterministic curve γ corresponds to a deterministic driving term $w(t)$, a random γ corresponds to a random $w(t)$. One obtains SLE(κ) by taking $w(t)$ to be a Brownian motion with speed κ:

Definition 5. *Schramm–Loewner evolution*, or SLE(κ), is the Loewner chain one obtains by taking $w(t) = \sqrt{\kappa} B_t$, $\kappa \in [0, \infty)$. Here B_t denotes the standard (speed one) Brownian motion (Wiener process).

The resulting slit will be almost surely a continuous curve. So we will also use the term SLE for the resulting random curve, i.e. a probability measure on the space of curves (to be rigorous one can think of a Borel measure on the space of curves with uniform norm). Different speeds κ produce different curves: we grow the slit with constant speed (measured by capacity), while the driving term "wiggles" faster. Naturally, the curves become more "fractal" as κ increases: for $\kappa \leq 4$ the curve is almost surely simple, for $4 < \kappa < 8$ it almost surely touches itself, and for $\kappa \geq 8$ it is almost surely space-filling (i.e. visits every point in $\mathbb{C}_+$) – see [22], [30] for these and other properties. Moreover, Vincent Beffara [7] has proved that the Hausdorff dimension of the SLE(κ) curve is almost surely $\min(1 + \kappa/8, 2)$.

3.3. Conformal Markov property. Suppose we want to describe the scaling limits of cluster perimeters, or interfaces for lattice models assuming their existence and conformal invariance. We follow Oded Schramm [32] to show that Brownian motion as the driving force arises naturally. Consider a simply connected domain Ω with two boundary points, a and b. Superimpose a lattice with mesh ε and consider some lattice model, say critical percolation with the Dobrushin boundary conditions, leading to an interface running from a to b, which is illustrated by Figure 1 for a rectangle with two opposite corners as a and b. So we end up with a random simple curve (a broken line) connecting a to b inside Ω. The law of the curve depends of course on the lattice superimposed. If we believe the physicists' predictions, as mesh tends to zero, this measure on broken lines converges (in an appropriate weak-$*$ topology) to some measure $\mu = \mu(\Omega, a, b)$ on continuous curves from a to b inside Ω.

In this setup the conformal invariance prediction can be formulated as follows:

(A) Conformal invariance. *For a conformal map ϕ of the domain Ω one has*

$$\phi\left(\mu(\Omega, a, b)\right) = \mu(\phi(\Omega), \phi(a), \phi(b)).$$

Here a bijective map $\phi : \Omega \to \phi(\Omega)$ induces a map acting on the curves in Ω, which in turn induces a map on the probability measures on the space of such curves, which we denote by the same letter. By a conformal map we understand a bijection which locally preserves angles.

Moreover, if we start drawing the interface from the point a, we will be walking around the grey cluster following the right-hand rule – see Figure 1. If we stop at some point a' after drawing the part γ' of the interface, we cannot distinguish the boundary of Ω from the part of the interface we have drawn: they both are colored grey on the

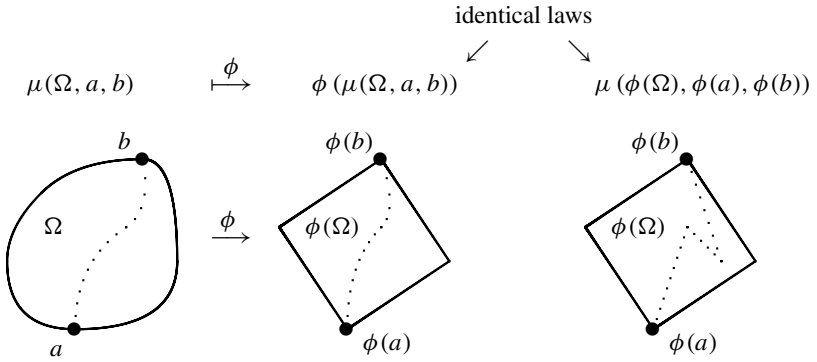


Figure 5. (A) Conformal invariance: conformal image of the law of the curve γ (dotted) in Ω coincides with the law of the curve γ in the image domain $\phi(\Omega)$.

(counterclockwise) arc $a'b$ and white on the arc ba' of the domain $\Omega \setminus \gamma'$. So we can say that the conditional law of the interface (conditioned on it starting as γ') is the same as the law in a new domain with a slit. We expect the limit law μ to have the same property:

(B) Markov property. *The law conditioned on the interface already drawn is the same as the law in the slit domain:*

$$\mu\,(\Omega, a, b)\,|\gamma' = \mu(\Omega \setminus \gamma', a', b).$$

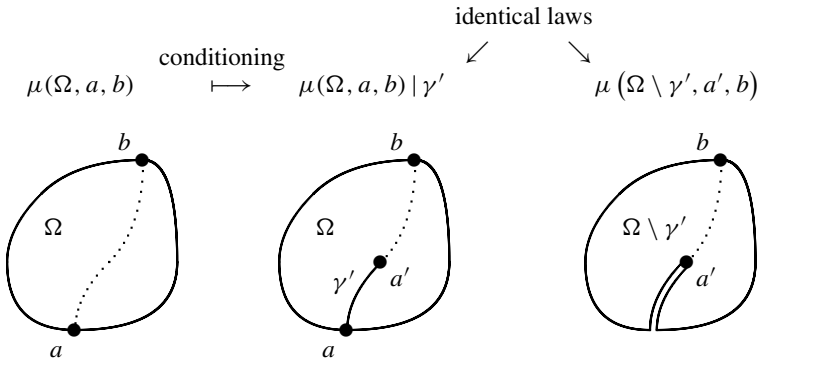


Figure 6. (B) Markov property: The law conditioned on the curve already drawn is the same as the law in the slit domain. In other words when drawing the curve we do not distinguish its past from the boundary.

If one wants to utilize these properties to characterize μ, by (A) it is sufficient to study some reference domain (to which all others can be conformally mapped), say

the upper half-plane $\mathbb{C}_+$ with a curve running from 0 to ∞. Given (A), the second property (B) is easily seen to be equivalent to the following:

(B′) Conformal Markov property. *The law conditioned on the interface already drawn is a conformal image of the original law. Namely, for any conformal map $G = G_{\gamma'}$ from $\mathbb{C}_+ \setminus \gamma'$ to $\mathbb{C}_+$ preserving ∞ and sending the tip of γ' to 0, we have*

$$\mu\,(\mathbb{C}_+, 0, \infty)\,|\gamma' = G^{-1}(\mu\,(\mathbb{C}_+, 0, \infty)).$$

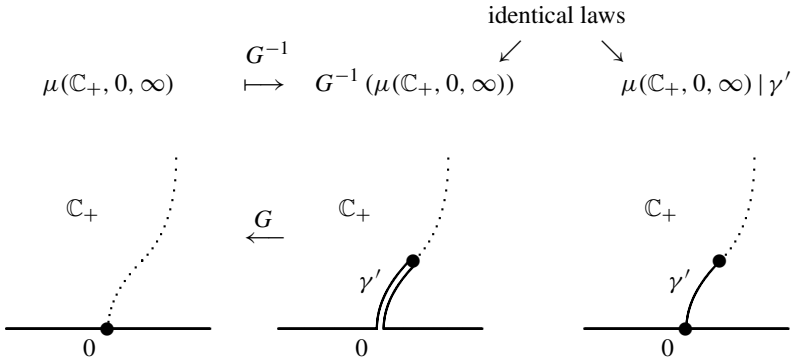


Figure 7. (B′) Conformal Markov property: The law conditioned on the curve already drawn is a conformal image of the original law. In other words the curve has "Markov property in conformal coordinates".

Remark 6. Note that property (B′) is formulated for the law μ for one domain only, say $\mathbb{C}_+$ as above. If we extend μ to other domains by conformal maps, it turns out that (A) and (B) are equivalent to (B′) together with scale invariance (under maps $z \mapsto kz, k > 0$).

To use the property (B′), we describe the random curve by the Loewner evolution with a certain random driving force $w(t)$ (we assume that the curve is almost surely an allowed slit). If we fix the time t, the property (B′) with the slit $\gamma[0, t]$ and the map $G_t(z) = g_t(z) - w(t)$ can be rewritten for random conformal map $G_{t+\delta}$ conditioned on G_t (which is the same as conditioning on $\gamma[0, t]$) as

$$G_{t+\delta}|G_t = G_t(G_\delta).$$

Expanding G's near infinity we obtain

$$\begin{aligned} z - w(t+\delta) + \cdots |G_t &= (z - w(t) + \cdots) \circ (z - w(\delta) + \cdots) \\ &= z - (w(t) + w(\delta)) + \cdots, \end{aligned}$$

concluding that

$$w(t+\delta) - w(t)|G_t = w(\delta).$$

This means that $w(t)$ is a continuous (by Loewner's theorem) stochastic process with independent stationary increments. Thus for *a random curve satisfying* (B′) the driving force $w(t)$ has to be a Brownian motion with a certain speed $\kappa \in [0, \infty)$ and drift $\alpha \in \mathbb{R}$:

$$w(t) = \sqrt{\kappa} B_t + \alpha t.$$

Applying (A) with anti-conformal reflection $\phi(u + iv) = -u + iv$ or with stretching $\phi(z) = 2z$ shows that α vanishes. So one logically arrives at the definition of SLE and the following

Schramm's principle. *A random curve satisfies* (A) *and* (B) *if and only if it is given by* SLE(κ) *for some* $\kappa \in [0, \infty)$.

The discussion above is essentially contained in Oded Schramm's paper [32] for the radial version, when slit is growing towards a point inside and the Loewner differential equation takes a slightly different form. To make this principle a rigorous statement, one has to require the curve to be almost surely an allowed slit.

4. SLE as a scaling limit

4.1. Strategy. In order to use the above principle one still has to show the *existence* and *conformal invariance* of the scaling limit, and then calculate some observable to pin down the value of κ. For percolation one can employ its locality or Cardy's formula for crossing probabilities to show that $\kappa = 6$. Based on this observation Oded Schramm concluded in [32] that *if* percolation interface has a conformally invariant scaling limit, it must be SLE(6).

But it is probably difficult to show that some interface has a conformally invariant scaling limit without actually identifying the latter.

To identify a random curve in principle one needs "infinitely many observables," e.g. knowing for any finite number of points the probability of passing above them. This seems to be a difficult task, which is doable for percolation since the locality allows us to create many observables from just one (crossing probability), see [38].

Fortunately it turns out that even in the general case if an observable has a limit satisfying analogues of (A) and (B), one can deduce convergence to SLE(κ) (with κ determined by the values of the observable).

This was demonstrated by Greg Lawler, Oded Schramm and Wendelin Werner in [23] in establishing the convergence of two related models: of loop erased random walk to SLE(2) and of uniform spanning tree to SLE(8).

They described the discrete curve by a Loewner evolution with unknown random driving force. Stopping the evolution at times t and s and comparing the values of the observable, one deduces (approximate) formulae for the conditional expectation and variance of the increments of the driving force. Skorokhod embedding theorem is then used to show that driving force converges to the Brownian motion. Finally

one has to prove a (stronger) convergence of the measures on curves. The trick is that knowing just one observable (but *for all domains*) after conditioning translates to a continuum of information about the driving force.

We describe a different approach with the same general idea, which is perhaps more transparent, separating "exact calculations" from "a priori estimates". The idea is to first get a priori estimates, which imply that collection of laws is precompact in a suitable space of allowed slits. Then to establish the convergence it is enough to show that limit of any converging subsequence coincides with SLE. To do that we describe the subsequential limit by Loewner evolution (with unknown random driving force $w(t)$) and extract from the observable enough information to evaluate expectation and quadratic variation of increments of $w(t)$. Lévy's characterization implies that $w(t)$ is the Brownian motion with a particular speed κ and so our curves converge to SLE(κ).

As an example we discuss below an alternative proof of convergence to SLE(6) in the case of percolation, which uses crossing probability as an observable. For a domain Ω with boundary points a, b superimpose triangular lattice with mesh ε and Dobrushin boundary conditions. We obtain an interface γ_ε (between open vertices on one side and closed on another) running from a to b, see Figure 1, i.e. a measure μ_ε on random curves running from a to b.

4.2. Compactness. First we note that the collection $\{\mu_\varepsilon\}$ is precompact (in weak-$*$ topology) in the space of continuous curves that are Loewner allowed slits.

The necessary framework for precompactness in the space of continuous curves was suggested by Michael Aizenman and Almut Burchard [2]. It turns out that appropriate bounds for probability of an annulus being traversed k times imply tightness: a curve has a Hölder parameterization with stochastically bounded norm. Hence $\{\mu_\varepsilon\}$ is precompact by Prokhorov's theorem: a (uniformly controlled) part of μ_ε is supported on a compact set (of curves with norm bounded by M), and so such parts are weakly precompact by Banach–Alaoglu theorem, whereas the mass of the remainder tends uniformly to zero as $M \to \infty$.

The curves on the lattice are simple, so they cannot have transversal self-intersections even after passing to the limit. So to check that for any weak limit of μ_ε's almost every curve is an allowed slit, one has to check that as we grow it the tip is always visible and moves continuously when viewed from infinity. Essentially, one has to rule out two scenarios: that the curve passes for a while inside already visited set, and that the curve closes a loop, and then travels inside before exiting. Both are reduced to probabilities of annuli traversing.

In the case of percolation one uses the Russo–Seymour–Welsh theory [31], [36] together with Michael Aizenman's observation [1] (that in the limit interface can visit no point thrice – "no 6 arms") to obtain the required estimates.

Since the collection of interface laws $\{\mu_\varepsilon\}$ is precompact (in weak-$*$ topology) in the space of continuous curves that are Loewner allowed slits, to show that as mesh goes to zero the interface law converge to the law of SLE(6), it is sufficient to show

that the limit of any converging subsequence is in fact SLE(6).

Take some subsequence converging to a random curve in the domain Ω from a to b. We map conformally to a half-plane $\mathbb{C}_+$, obtaining a curve γ from 0 to ∞ with law μ. We must show that μ is given by SLE(6).

By a priori estimates γ is almost surely an allowed slit. So we can describe γ by a Loewner evolution with a (random) driving force $w(t)$. It remains to show that $w(t) = \sqrt{6}B_t$. Note that at this point we only know that $w(t)$ is an almost surely continuous random function – we do not even have a Markov property.

4.3. Martingale observable. Given a topological rectangle (a simply connected domain Ω with boundary points a, b, c, d) one can superimpose a lattice with mesh ε onto Ω and study the probability $\Pi_\varepsilon(\Omega, [a, b], [c, d])$ that there is an open cluster joining the arc $[a, b]$ to the arc $[c, d]$ on the boundary of Ω. It is conjectured that there is a limit $\Pi := \lim_{\varepsilon\to 0} \Pi_\varepsilon$, which is conformally invariant (depends only on the conformal modulus of the configuration Ω, a, b, c, d), and satisfies Cardy's formula (predicted by John Cardy in [9] and proved in [37]) in half-plane:

$$\Pi(\mathbb{C}_+, [1-u, 1], [\infty, 0]) = \frac{\Gamma(2/3)}{\Gamma(1/3)\Gamma(4/3)}\, u^{1/3}\, {}_2F_1\left(\frac{1}{3}, \frac{2}{3}; \frac{4}{3}; u\right) =: F(u). \quad (5)$$

Above ${}_2F_1$ is the hypergeometric function, so one can alternatively write

$$F(u) = \int_0^u (v(1-v))^{-2/3}\, dv \Big/ \int_0^1 (v(1-v))^{-2/3}\, dv.$$

Particular nature of the function is not important, we rather use the fact that there is an explicit formula for half-plane with four marked boundary points and hence by conformal invariance for an arbitrary topological rectangle. The value $\kappa = 6$ will arise later from some expression involving derivatives of F.

Assume that for some percolation model we are able to prove the above conjecture (for critical site percolation on the triangular lattice it was proved in [37], [38]).

Add two points on the boundary, making Ω a topological rectangle $axby$ and consider the crossing probability $\Pi_\varepsilon(\Omega, [a, x], [b, y])$ (from the arc ax to the arc by on a lattice with mesh ε).

Parameterize the interface γ_ε in some way by time, and draw the part $\gamma_\varepsilon[0, t]$. Note that it has open vertices on one side (arc $\gamma_\varepsilon(t)a$) and closed on another (arc $a\gamma_\varepsilon(t)$). Then any open crossing from the arc by to the arc ax inside Ω is either disjoint from $\gamma_\varepsilon[0, t]$, or hits its "open" arc $\gamma_\varepsilon(t)a$. In either case it produces an open crossing from the arc by to the arc $\gamma_\varepsilon(t)x$ inside $\Omega \setminus \gamma_\varepsilon[0, t]$, and converse also holds. Therefore one sees that for every realization of $\gamma_\varepsilon[0, t]$ the crossing probability conditioned on $\gamma_\varepsilon[0, t]$ coincides with crossing probability in the slit domain $\Omega \setminus \gamma_\varepsilon[0, t]$:

$$\Pi_\varepsilon(\Omega, [a, x], [b, y] | \gamma_\varepsilon[0, t]) = \Pi_\varepsilon(\Omega \setminus \gamma_\varepsilon[0, t], [\gamma_\varepsilon(t), x], [b, y]), \quad (6)$$

an analogue of the Markov property (B). Alternatively this follows from the fact that Π can be understood in terms of the interface as the probability that it touches the

arc xb before the arc by. For example, in Figure 1 there is no horizontal grey (open) crossing (there is a vertical white crossing instead), and interface traced from the left upper corner a touches the lower side xb before the right side by.

Stopping the curve at times $t < s$ and using (6) we can write by the total probability theorem for every realization of $\gamma_\varepsilon[0,t]$

$$\begin{aligned}&\Pi_\varepsilon\left(\Omega \setminus \gamma_\varepsilon[0,t], [\gamma_\varepsilon(t), x], [b, y]\right)\\ &\quad = \mathbb{E}_{\gamma_\varepsilon[t,s]}\left(\Pi_\varepsilon\left(\Omega \setminus \gamma_\varepsilon[0,s], [\gamma_\varepsilon(s), x], [b, y]\right) | \gamma[0,t]\right).\end{aligned} \tag{7}$$

The same a priori estimates as in the previous subsection show that the identity (7) also holds for the (subsequential) scaling limit μ (strictly speaking there is an error term in case the interface touches the arcs $[ax]$ or $[ya]$ before time s, but it decays very fast as we move x and y away from a). We know that the scaling limit $\Pi := \lim_{\varepsilon\to 0} \Pi_\varepsilon$ of the crossing probabilities exists and is conformally invariant, so we can rewrite (7) for the curve γ with Loewner parameterization as

$$\begin{aligned}&\Pi\left(\mathbb{C}_+ \setminus \gamma[0,t], [\gamma(t), x], [\infty, y]\right)\\ &\quad = \mathbb{E}_{\gamma[t,s]}\left(\Pi\left(\mathbb{C}_+ \setminus \gamma[0,s], [\gamma(s), x], [\infty, y]\right) | \gamma[0,t]\right),\end{aligned} \tag{8}$$

for almost every realization of $\gamma[0,t]$. Moreover we can plug in exact values of the crossing probabilities, given by the Cardy's formula. Recall that the domain $\mathbb{C}_+ \setminus \gamma[0,t]$ is mapped to half-plane by the map $g_t(z)$ with $\gamma(t) \mapsto w(t)$. Then the map $z \mapsto \frac{g_t(z)-g_t(y)}{g_t(x)-g_t(y)}$ also maps it to half-plane with $\gamma(t) \mapsto \frac{w(t)-g_t(y)}{g_t(x)-g_t(y)}$, $y \mapsto 0$, $x \mapsto 1$. Using conformal invariance and applying Cardy's formula we write

$$\begin{aligned}\Pi\left(\mathbb{C}_+ \setminus \gamma[0,t], [\gamma(t), x], [\infty, y]\right) &= \Pi\Big(\mathbb{C}_+, \Big[-\frac{g_t(y)-w(t)}{g_t(x)-g_t(y)}, 1\Big], [\infty, 0]\Big)\\ &= F\Big(\frac{g_t(x)-w(t)}{g_t(x)-g_t(y)}\Big),\end{aligned} \tag{9}$$

for Cardy's hypergeometric function F.

4.4. Conformally invariant martingale. Plugging (9) into both sides of (8) we arrive at

$$F\Big(\frac{g_t(x)-w(t)}{g_t(x)-g_t(y)}\Big) = \mathbb{E}_{\gamma[t,s]}\Big(F\Big(\frac{g_s(x)-w(s)}{g_s(x)-g_s(y)}\Big) | \gamma[0,t]\Big). \tag{10}$$

Remark 7. Denote by $x_t := g_t(x) - w(t)$ and $y_t := g_t(y) - w(t)$ trajectories of x and y under the random Loewner flow. Then (10) essentially means that $F\big(\frac{x_t}{x_t-y_t}\big)$ is a martingale.

Since we want to extract the information about $w(t)$, we fix the ratio $x/(x-y) := 1/3$ (anything not equal to $1/2$ would do) and let x tend to infinity: $y := -2x$, $x \to +\infty$. Using the normalization $g_t(z) = z + 2t/z + \mathcal{O}(1/z^2)$ at infinity, writing

Taylor expansion for F, and plugging in values of derivatives of F at $1/3$, we obtain the following expansion for the right-hand side of (10):

$$\begin{aligned}
\cdots &= F\left(\frac{x - w(t) + 2t/x + \mathcal{O}(1/x^2)}{(x + 2t/x + \mathcal{O}(1/x^2)) - (-2x + 2t/(-2x) + \mathcal{O}(1/x^2))}\right) \\
&= F\left(\frac{1}{3} - \frac{w(t)}{3}\frac{1}{x} + \frac{t}{3}\frac{1}{x^2} + \mathcal{O}\left(\frac{1}{x^3}\right)\right) \\
&= F\left(\frac{1}{3}\right) - \frac{w(t)}{3}F'\left(\frac{1}{3}\right)\frac{1}{x} + \left(\frac{t}{3}F'\left(\frac{1}{3}\right) + \frac{w(t)^2}{3^2\cdot 2}F''\left(\frac{1}{3}\right)\right)\frac{1}{x^2} + \mathcal{O}\left(\frac{1}{x^3}\right) \\
&= F\left(\frac{1}{3}\right) - \frac{1}{x}\frac{\Gamma(2/3)}{\Gamma(1/3)\Gamma(4/3)}\frac{3^{1/3}}{2^{2/3}}\mathbb{E}\, w(t) \\
&\qquad - \frac{1}{x^2}\frac{\Gamma(2/3)}{\Gamma(1/3)\Gamma(4/3)}\frac{1}{3^{2/3}2^{5/3}}\mathbb{E}\left(w(t)^2 - 6t\right) + \mathcal{O}\left(\frac{1}{x^3}\right) \\
&=: A - \frac{1}{x}B\,\mathbb{E}\, w(t) - \frac{1}{x^2}C\,\mathbb{E}\left(w(t)^2 - 6t\right) + \mathcal{O}\left(\frac{1}{x^3}\right),
\end{aligned}$$

where we plugged in values of the derivative for hypergeometric function. Using similar reasoning for the right-hand side of (10) we arrive at the following identity:

$$\begin{aligned}
&A - \frac{1}{x}B\,\mathbb{E}\, w(t) - \frac{1}{x^2}C\,\mathbb{E}\left(w(t)^2 - 6t\right) + \mathcal{O}\left(\frac{1}{x^3}\right) \\
&\quad = A - \frac{1}{x}B\,\mathbb{E}_{\gamma[t,s]}\left(w(s)|\gamma[0,t]\right) - \frac{1}{x^2}C\,\mathbb{E}_{\gamma[t,s]}\left(w(s)^2 - 6s|\gamma[0,t]\right) + \mathcal{O}\left(\frac{1}{x^3}\right).
\end{aligned}$$

Equating coefficients in the series above, we conclude that

$$\mathbb{E}_{w[t,s]}\left(w(s)|w[0,t]\right) = 0, \quad \mathbb{E}_{w[t,s]}\left(w(s)^2 - 6s|w[0,t]\right) = w(t)^2 - 6t. \tag{11}$$

Thus $w(t)$ is a continuous (by Loewner's theorem) process such that both

$$w(t) \quad \text{and} \quad w(t)^2 - 6t$$

are martingales so by Lévy's characterization of the Brownian motion $w(t) = \sqrt{6}B_t$, and therefore SLE(6) is the scaling limit of the critical percolation interface.

The argument will work wherever Cardy's formula and a priori estimates are available, particularly for triangular lattice. More generally, any conformally invariant martingale will do, with value of κ arising from its Taylor expansion.

Remark 8. The scheme can also be reversed to do calculations for SLE's, if an observable is a martingale (e.g. crossing probability). Indeed, writing the same formulae with $x/(x-y) = a$ we conclude that the coefficient by $\frac{1}{x^2}$, namely

$$\frac{2a(1-2a)}{1-a}tF'(a) + \frac{a}{2}\mathbb{E}\left(w(t)^2\right)F''(a)$$

vanishes. Since for $w(t) = \sqrt{6}B(t)$ one has $\mathbb{E}\left(w(t)^2\right) = 6t$, we arrive at the differential equation

$$\frac{2(1-2a)}{3(1-a)}F'(a) + F''(a) = 0.$$

With the given boundary data it has a unique solution, which is Cardy's hypergeometric function.

5. Ising model and beyond

The martingale method as described above shows that to construct a conformally invariant scaling limit for some model we need a priori estimates and a non-trivial martingale observable with a conformally invariant scaling limit.

5.1. A priori estimates. A priori estimates are necessary to show that collection of interface laws is precompact in weak-$*$ topology (on the space of measures on continuous curves which are allowed slits).

If we follow the same route as for percolation (via the work [2] of Michael Aizenman and Almut Burchard), we only need to evaluate probabilities of traversals of an annulus in terms of its modulus. For percolation such estimates are (almost) readily available from the Russo–Seymour–Welsh theory. For uniform spanning tree and loop erased random walk one can derive the estimates using random walk connection and the known estimates for the latter (a "branch" of a uniform spanning tree is a loop erased random walk), see [3], [32].

For the Ising model the required estimates do not seem to be readily available, but a vast arsenal of methods is at hand. Essentially all we need can be reduced by monotonicity arguments to spin correlation estimates of Bruria Kaufman, Lars Onsager and Chen Ning Yang [14], [44].

For general random cluster or loop models such exact results are not available, but we actually need much weaker statements, and many of the techniques used by us for the Ising model (like FKG inequalities) are well-known in the general case.

So this part does not seem to be the main obstacle to construction of scaling limits, though it might require very hard work. Moreover, following the proposed approach we actually get that interfaces have a Hölder parameterization with uniformly stochastically bounded norm. Thus rather weak kinds of convergence of interfaces would lead to convergence in uniform norm (or rather weak-$*$ convergence of measures on curves with uniform norm).

It also appears that the same a priori estimates can be employed to show observable convergence in the cases concerned, and hopefully they will be sufficient for other models. So a more pressing question is how to construct a martingale observable.

5.2. Conformally covariant martingales. Suppose that for every simply connected domain Ω with a boundary point a we have defined a random curve γ starting from a.

Mark several points $b, c, \dots$ in Ω or on the boundary. Remark 7 suggests the following definition:

Definition 9. We say that a function (or rather a *differential*) $F(\Omega, a, b, c, \dots)$ is a *conformal* (*covariant*) *martingale* for a random curve γ if

F is conformally covariant:

$$F(\Omega, a, b, c, \dots) = F(\phi(\Omega), \phi(a), \phi(b), \phi(c), \dots) \cdot \phi'(b)^{\alpha} \bar{\phi}'(b)^{\beta} \phi'(c)^{\gamma} \bar{\phi}'(c)^{\delta} \dots, \tag{12}$$

and

$$F(\Omega \setminus \gamma[0, t], \gamma(t), b, c, \dots) \text{ is a martingale} \tag{13}$$

with respect to the random curve γ drawn from a (with Loewner parameterization).

Introducing covariance at $b, c, \dots$ we do not ask for covariance at a, since it always can be rewritten as covariance at other points. And applying factor at a would be troublesome: once we started drawing a curve the domain becomes non-smooth in its neighborhood, creating problems with the definition.

If the exponents $\alpha, \beta, \dots$ vanish, we obtain an invariant quantity. While the crossing probability for the percolation was invariant, many quantities of interest in physics are covariant differentials, e.g. open edge density at c would scale as a lattice step to some power (depending on the model), so we would arrive at a factor

$$|\phi'(c)|^{\delta} = \phi'(c)^{\delta/2} \bar{\phi}'(c)^{\delta/2}.$$

There are other possible generalisations, e.g. one can add the Schwarzian derivative of ϕ to (12).

The two properties in Definition 9 are analogues of (A) and (B), and similarly combined they show that for the curve γ mapped to half-plane from any domain Ω so that $a \mapsto 0$, $b \mapsto \infty$, $c \mapsto x$ (note that the image curve in $\mathbb{C}_+$ might depend on Ω – we only know the conformal invariance of an observable, not of the curve itself) we have an analogue of (B′), which was already mentioned in Remark 7 for percolation. Namely

$$F(\mathbb{C}_+, 0, \infty, g_t(x), \dots) \cdot g_t'(x)^{\gamma} \bar{g}_t'(x)^{\delta} \dots,$$

is a martingale with respect to the random Loewner evolution (covariance factor at $b = \infty$ is absent, since $g_t'(\infty) = 1$).

The equation (10) can be written for this F, and if we can evaluate F exactly, the same machinery as one used by Greg Lawler, Oded Schramm and Wendelin Werner in [23] or as the one discussed above for percolation proves that our random curve is SLE. So one arrives at a following generalization of Oded Schramm's principle:

Martingale principle. *If a random curve γ admits a* (*non-trivial*) *conformal martingale F, then γ is given by* SLE *with κ* (*and drift depending on modulus of the configuration*) *derived from F.*

Remark 10. In chordal situation we consider curves growing from a towards another boundary point b in a simply connected domain. But the same conclusion would hold on general domains or Riemann surfaces with boundary once we find a covariant martingale (for appropriate generalizations of Loewner evolutions see e.g. the book [4]). The only difference is that driving force of the corresponding Loewner evolution will be a Brownian motion with drift depending on conformal modulus of the configuration $\Omega, a, b, c, \dots$, leading to SLE generalizations. Starting from lattice models with various boundary conditions and conditioned on various events, one can see which drifts will be of interest for SLE generalizations.

5.3. Discrete analyticity. Passing to the lattice model, we want to find a discrete object, which in the limit becomes a conformally covariant martingale.

Martingale property is actually more accessible in the discrete setting. For example, functions which are defined as observables (like probability of the interface going through a vertex, edge density for the model, etc.) have the martingale property built in, and so only conformal covariance must be established.

Alternatively, one can work with a discrete function $F(\Omega, a, b, c)$ (a priori not related to lattice models) which has a conformally covariant scaling limit by construction. Then we need to connect it to a particular lattice model, establishing a martingale property (13). In the discrete case it is sufficient to check the latter for a curve advanced by one step. Assume that once we have drawn the part γ' of the interface from the point a to point a', it turns left with probability $p = p(\Omega, \gamma', a', b)$ creating a curve $\gamma_l = \gamma \cup \{a_l\}$ or right with probability $(1 - p)$ creating a curve $\gamma_r = \gamma \cup \{a_r\}$. Then it is enough to check the identity

$$F(\Omega \setminus \gamma', a', b, c) = pF(\Omega \setminus \gamma_l, a_l, b, c) + (1 - p)F(\Omega \setminus \gamma_r, a_r, b, c), \qquad (14)$$
$$p = p(\Omega, \gamma', a', b).$$

Actually our proof for the Ising model can be rewritten that way, with F defined as a solution of an appropriate discretization of the Riemann Boundary Value Problem (17) – the observable nature of F never comes up.

Moreover, starting with F one can define a random curve by choosing "turning probabilities" p so that identity (14) is satisfied, obtaining a model with conformally invariant scaling limit by "reverse engineering." For example, starting with a harmonic function of c with boundary values 1 on the arc ba and 0 on the arc ab, one obtains a unique discrete random curve, which has it as a martingale. Note that such a function is a particular case $\alpha = 1$ of the martingale (15) below, corresponding to $\kappa = 4$ (or rather its integral). In [35] Oded Schramm and Scott Sheffield introduced this curve with a nicer "Harmonic Explorer" definition, and utilizing the mentioned observable showed that it indeed converges to SLE(4). It seems that in this way one can use the solutions to the problem (17) to construct models converging to arbitrary SLE's, however it is not clear though whether they would similarly have "nicer" definitions.

Anyway, for either approach to work we need a *discrete conformal covariant* with a scaling limit. We have tried discretizations of many conformally invariant

objects (extremal length, capacity, solutions to variational problems, …) and the most promising in this context seem to be discrete harmonic or analytic functions in additional variable(s) (in $c, \dots$). Firstly, all other invariants can be rewritten in this way. Secondly, discretization of harmonic and analytic functions is a nice and very well studied (especially in the case of harmonic ones) object. Thirdly, one can obtain very non-trivial invariants by just checking local conditions: harmonicity or analyticity inside plus some boundary conditions (Dirichlet, Neumann, Riemann–Hilbert, etc.). The most natural candidate would be a harmonic function solving some Dirichlet problem.

Note that such an observable is known for the Brownian motion. A classical theorem [13] of Shizuo Kakutani states that in a domain Ω exit probabilities for Brownian motion started at z are harmonic functions in z with easily determinable boundary values. Though Kakutani works directly with Brownian motion, one can do the same for the random walk (which is actually much easier, since discrete Laplacian of the exit probability is trivially zero), and then passing to a limit deduce statements about Brownian motion, including its conformal invariance.

5.4. Classification of conformal martingales. Before we start working in the discrete setting, we might want to investigate which functions are conformal martingales for SLE curves, and so can arise as scaling limits of martingale observables for lattice models.

As discussed in Remark 8, one can write partial differential equations for SLE conformal martingales. For small number of points those equations can be solved, and in such a way one computes dimensions, scaling exponents and other quantities of interest. For any particular value of κ we can see which martingales have the simplest form and so are probably easier to work with. Also if they have a geometric SLE interpretation (like probability of SLE curve going to one side of a point, etc.) we can study similar quantities for the lattice model.

It turns out that only for $\kappa = 4$ one obtains a nice harmonic martingale with Dirichlet boundary conditions. In that case the probability of SLE(4) passing to one side of a point z is harmonic in z and has boundary conditions 0 and 1, see [33]. Oded Schramm and Scott Sheffield [35] constructed a model which has this property on discrete level built in. Unfortunately the property was not yet observed in any of the classical models conjecturally converging to SLE(4), though results of Kenyon [16] show it holds for double-domino curves in Temperley domains (i.e. a domain with the boundary satisfying a certain local condition).

In the case of mixed Dirichlet–Neumann conditions, it becomes possible to work with some other values of κ, including uniform spanning tree $\kappa = 8$, which is exploited in [23]. There are also covariant candidates for a few other values of κ (notably $8/3$ which corresponds to self-avoiding random walk), but they were not yet observed in lattice models.

Thus to study general models, one is forced to utilize more general boundary value problems with a Riemann(–Hilbert) Boundary Value Problem being the natural

candidate. Besides harmonic function it involves its harmonic conjugate, and so is better formulated in terms of analytic functions. Moreover, discrete analyticity involves a first order Cauchy–Riemann operator, rather than a second order Laplacian, and so it should be easier to deal with than harmonicity.

As discussed above, we can classify all analytic martingales. For chordal SLE and F with three points a, b, z as parameters we discover two particularly nice families. The following proposition will be discussed in [39] and our subsequent work:

Proposition 11. *Let Ω be a simply connected domain with boundary points a, b. Let $\Phi(z) = \Phi(\Omega, a, b, z)$ be a mapping of Ω to a horizontal strip $\mathbb{R} \times [0, 1]$, such that a and b are mapped to $\mp\infty$. Then*

$$F(\Omega, a, b, z) = \Phi'(z)^\alpha \quad \text{with } \alpha = \frac{8}{\kappa} - 1, \tag{15}$$

is a martingale for SLE(κ). *Let $\Psi(z) = \Psi(\Omega, a, b, z)$ be a mapping of Ω to a half-plane $\mathbb{C}_+$, such that a and b are mapped to ∞ and 0 correspondingly. Then*

$$G(\Omega, a, b, z) = \Psi'(z)^\alpha \Psi'(b)^{-\alpha} \quad \text{with } \alpha = \frac{3}{\kappa} - \frac{1}{2}, \tag{16}$$

is a martingale for SLE(κ).

These martingales make most sense for $\kappa \in [4, 8]$ and $\kappa \in [8/3, 8]$ correspondingly, and are related to observables of interest in Conformal Field Theory (which was part of our motivation to introduce them). Note that both functions are covariant with power α (which is the spin in physics terminology), and solve the Riemann boundary value problems

$$\operatorname{Im}\left(F(z)\tau(z)^\alpha\right) = 0, \quad z \in \partial\Omega, \tag{17}$$

where $\tau(z)$ is the tangent vector to $\partial\Omega$ at z.

The problem is to observe these functions in the discrete setting, and some intuition can be obtained from their geometric meaning for SLE's. For example, F is roughly speaking (one has to consider an intermediate scale to make sense of it) an expectation of SLE curve passing through z taken with some complex weight depending on the winding.

5.5. Height models and Coulomb gas. The above-mentioned expectation actually makes more sense (and is immediately well-defined) in the discrete setting and one arrives at the same object with the same complex weight via several different approaches.

One way is to consider the Coulomb gas arguments (cf. [29] by Bernard Nienhuis) for the loop representation. In the random cluster case at criticality, the weight of a loop is $\sqrt{q}$ – recall (2). We randomly and independently orient the loops, and introduce the height function h which whenever a loop is crossed changes by ± 1 (depending on loop direction – think of a topographic map). One could weight oriented loops by

$\sqrt{q}/2$, obtaining essentially the same model. However it makes sense to consider a complex weight instead. When q is in the $[0, 4]$ range, there is a complex unit number $\mu = \exp(k \cdot 2\pi i)$ such that

$$k = \frac{1}{2\pi}\arccos\left(\sqrt{q}/2\right) \text{ or } \mu + \bar{\mu} = \sqrt{q}. \tag{18}$$

We (independently and randomly) orient all loops, prescribing weight μ per counter-clockwise and $\bar{\mu}$ per clockwise loop.

Forgetting orientation of loops reconstructs the original model. Unfortunately the new partition function is complex and no longer leads to a probability measure (moreover, its variation blows up as the lattice step goes to zero), but it can be defined locally, making it much more accessible.

Indeed, going around a cycle, and turning by Δ_z at vertex z, the total sum of turns $\sum_{z\in\text{cycle}} \Delta_z$ is $\pm 2\pi$ depending on whether the cycle is counter or clockwise. So the weight per cycle can be written as $\prod_{z\in\text{cycle}} \exp(ik \cdot \Delta_z)$ and the total weight of the configuration is $\prod_{z\in\Omega} \exp(ik \cdot \Delta_z)$, which can be computed locally (without reference to the global order of cycles). The same weight can also be written in terms of the gradient of height function.

The interface is always oriented from a to b, so that the height function is always equal to 0 on the arc ab and to 1 on the arc ba. From physics arguments the interface curve (being "attached" to the boundary on both sides) should be weighted differently from loops, namely by $\exp(i(2k - 1/2) \cdot \Delta_z)$ per turn. When interface runs between two boundary points (being oriented from a to b), these factors do not matter, since total turn from a to b is independent of the configuration.

However, if we choose a point z on an interface and reverse the orientation of one of its halves (so that it is oriented from a to z and from z to b), the interface inputs a non-constant complex factor. This orientation reversal has a nice meaning: after it the height function acquires a $+2$ monodromy at z: when we go around z we cross two curves (halves of the interface) incoming into it.

All the loops (when we forget their orientation) still contribute the same $\sqrt{q}$ per loop, and the complex weight can be expressed in terms of the interface winding (total turn expressed in radians) from b to z, denoted by $w(\gamma, b \to z)$. So one logically arrives at the partition function Z for our model with $+2$ monodromy at z:

$$F(\Omega, a, b, z) := Z_{+2 \text{ monodromy at } z} = \mathbb{E}\chi_{z\in\gamma} \exp\left(i(4k - 1)w(\gamma, b \to z)\right). \tag{19}$$

This function is clearly a martingale, and there are strong indications (both from mathematics and physics points of view) that it is discrete analytic.

This follows from the fact that the interface can arrive at a boundary point z from b with a unique winding equal to the winding of the boundary from b to z, so we can express it in terms of the tangent vector $\tau(z)$. Writing this down, we discover that the function F solves a discrete version of the Riemann Boundary Value Problem (17) with $\alpha = 1 - 4k$.

Remark 12. The continuum problem was solved by the function (15), so if we establish discrete analyticity it only remains to show that a solution to a discrete Riemann Boundary Value Problem converges to its continuum counterpart. Moreover, combining identity $\alpha = 1 - 4k$ with (15) and (18) we obtain the relation between κ and q stated in Conjecture 4.

This convergence problem seems to be difficult (and open in the general case). The way we solve it in the Ising case is sketched below.

There are other indications that this function is nice to work with. Indeed, the easiest form of discrete analyticity involves local partial difference relations, and to prove those we should count configurations included into our expectation. To obtain relations, we need some bijections in the configuration space, and the easiest ones are given by local rearrangements (we worked with global rearrangements for percolation [37], but such work must be more difficult for non-local models).

The easiest rearrangement involves redirecting curves passing through z, see Figure 8, and we have a good control over relative weights of configurations whenever they are defined through windings. Counting how much a pair of configurations contributes to values of F at neighbors of z, we get some relations. Moreover, a careful analysis shows that the maximal number of relations is attained with the complex weight (19).

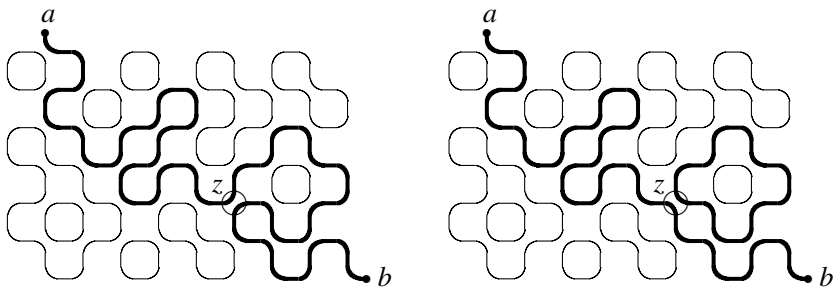


Figure 8. Rearrangement at a point z: we only change connections inside a small circle marking z. Either interface does not pass through z in both configurations, or it passes in a way similar to the pictured above. On the left the interface (in bold) passes through z twice, on the right (after the rearrangement) it passes once, but a new loop through z appears (also in bold). The loops not passing through z remain the same, so the weights of configurations differ by a factor of $\sqrt{q}$ because of the additional loop on the right. To get some linear relation on values of F, it is enough to check that any pair of such configurations makes equal contributions to two sides of the relation.

5.6. Ising model. We finish with a sketch of our proof for the random cluster representation of the Ising model (i.e. $q = 2$) on the square lattice $\varepsilon\mathbb{Z}^2$ at the critical temperature. As before consider loop representation in a simply connected domain Ω with two boundary points a and b and Dobrushin boundary conditions.

Consider function $F = F_\varepsilon(\Omega, a, b, z)$ given by (19) which is the expectation that interface from a to b passes through a vertex z taken with appropriate unit complex weight. Note that for Ising $q = 2$, so $k = 1/8$ and the weight is Fermionic (which of course was expected): a passage in the same direction but with a 2π twist has a relative weight -1, whereas a passage in the opposite direction with a counterclockwise π twist has a relative weight $-i$.

As discussed F automatically has the martingale property when we draw γ starting from a, so only conformal invariance in the limit has to be checked.

Color lattice vertices in chessboard fashion, and to each edge e prescribe orientation such that it points from a black vertex to a white one, turning it into a vector, or equivalently a complex number e. Denote by $\ell(e)$ the line passing through the origin and $\sqrt{\bar{e}}$ – the square root of the complex conjugate to e (the choice of the square root is not important). Careful analysis of the rearrangement in Figure 8 shows that F satisfies the following relation: for every edge $e \in \Omega$ orthogonal projections of the values of F at its endpoints on the line $\ell(e)$ coincide. We denote this common projection by $F(e)$ as it would also be given by the same formula (19) with z taken on the edge e (to be exact one has to divide by $2\cos(\pi/8)$ to arrive at the same normalization).

It turns out to be a form of *discrete analyticity*, and implies (but does not follow from) the common definition. The latter asks for the discrete version of the Cauchy–Riemann equations $\partial_{i\alpha} F = i\partial_\alpha F$ to be satisfied. Namely for every lattice square the values of F at four corners (denoted u, v, w, z in the counter-clockwise direction) should obey

$$F(z) - F(v) = i(F(w) - F(u)).$$

Remark 13. In the complex plane *holomorphic* (i.e. having a complex derivative) and *analytic* (i.e. admitting a power series expansion) functions are the same, so the terms are often interchanged. Though the term *discrete analytic* is in wide use, in discrete setting there are no power expansions, so it would be more appropriate to speak of *discrete holomorphic* (or *discrete regular*) functions.

As discussed above, F solves a discrete version of the Riemann Boundary Value Problem (17) with $\alpha = 1 - 4k = 1/2$, which was solved in the continuum case by $\sqrt{\Phi'}$. It remains to show that as the lattice step goes to zero, properly normalized F converges to the latter.

A logical thing to do is to integrate F^2 to retrieve Φ. Unfortunately, the square of a discrete analytic function is no longer discrete analytic and so cannot be integrated. However it turns out that there is a unique function $H = \operatorname{Im} \int F^2 dz$, which is defined on the dual lattice by

$$H(b) - H(w) = |F(e)|^2, \tag{20}$$

where edge e separates the centers of two adjacent squares, black b and white w.

After writing (20), one checks that

1. H is well defined and unique up to an additive constant,
2. H restricted to white (black) squares is super (sub) harmonic,

3. $H = 1$ on (counterclockwise) boundary arc ba and $H = 0$ on (counterclockwise) boundary arc ab,
4. The (local) difference between H restricted to white and to black squares tends uniformly to zero.

The properties 1, 2 are consequences of discrete analyticity: 1 a rather direct one, while 2 follows from the identity

$$\Delta H(u) = \pm |F(x) - F(y)|^2 ,$$

where u is a center of white (black) square with two opposite corner vertices x and y (particular choice is unimportant). Definition of F implies the property 3. The property 4 easily follows from a priori estimates (namely Kaufman–Onsager–Yang results [14], [44]). In principle it should also directly follow from the discrete analyticity of F and the property 3.

We immediately infer that H converges to $\mathrm{Im}\Phi$, and after differentiating and taking a square root we obtain the following:

Proposition 14. *Suppose that the lattice mesh ε_j goes to zero and a lattice domain Ω_j with boundary points a_j, b_j converges (in a weak sense, e.g. in Carathéodory metric) to a domain Ω with boundary points a, b as $j \to \infty$. Then away from the boundary there is a uniform convergence:*

$$\frac{1}{\sqrt{\varepsilon_j}} F(\Omega_j, a_j, b_j, z) \rightrightarrows \sqrt{\Phi'(\Omega, a, b, z)}$$

Since by Proposition 11 the function on the right is a martingale for SLE(16/3), convergence of the interface to the Schramm–Löwner evolution with $\kappa = 16/3$ follows.

6. Conclusion

At the moment the approach discussed above works only for a (finite) number of models. Another notable case when it works is the usual spin representation of the Ising model at critical temperature on the square lattice, pictured in Figure 2, where considering a similar observable (partition function with $+1$ monodromy, cf. (19)) leads to the martingale (16) and to Schramm–Loewner evolution with $\kappa = 3$. Interestingly, exactly the same definition of discrete analyticity arises.

Analogously, examination of partition function with $+1$ monodromy at z for hexagonal loop models (for all values of n at criticality) suggests its convergence to conformal martingale (16). These considerations lead to a new explanation of the Nienhuis' Conjecture 2 for the critical value of x. In this case we firmly believe that our method works all the way for $n = 1$ constructing conformally invariant scaling limits for the $O(1)$ model, but convergence estimates still have to be verified.

Two parallel methods, with observables related to (15) and (16), seem specially adapted to the square lattice and the hexagonal lattice correspondingly. However, the main arguments work for a large family of four- and trivalent graphs correspondingly. So we advance towards establishing the universality conjectures.

Though only for a few models the conformal invariance was proved, the only essential missing step for the remaining ones is discrete analyticity, and it can be attacked in a large number of ways.

So from our point of view, the perspectives for establishing conformal invariance of classical 2D lattice models are quite encouraging. Moreover, we can start discussing reasons for universality, and try to construct the full loop ensemble starting from the discrete picture. The approach discussed above is rigorous, but what makes it (and the whole SLE subject) even more interesting is that while borrowing some intuition from physics, it gives a new way to approach these phenomena.

Acknowledgments. Much of the work was completed while the author was a Royal Swedish Academy of Sciences Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation. The author also gratefully acknowledges support of the Swiss National Science Foundation.

Existence of a discrete analytic function in the Ising spin model which has potential to imply convergence of interfaces to SLE(3) was first noticed by Rick Kenyon and the author based on the dimer techniques applied to the Fisher lattice. However at the moment the Riemann Boundary Value Problem seemed beyond reach. John Cardy independently observed that (the classical version) of discrete analyticity holds for the function (19) restricted to edges.

I would like to thank Lennart Carleson for introducing me to this area, as well as for constant encouragement and advice. Most of what I know about lattice models was learnt from others, and I am especially grateful to Michael Aizenman, John Cardy and Rick Kenyon for numerous inspiring conversations on the subject. Much of the SLE considerations discussed above are due to Greg Lawler, Oded Schramm and Wendelin Werner. Finally I wish to thank Dmitri Beliaev, Ilia Binder and Geoffrey Grimmett for helpful comments on this note.

References

[1] Aizenman, M., The geometry of critical percolation and conformal invariance. In *STATPHYS 19* (Xiamen, 1995), World Scientific Publishing, River Edge, NJ, 1996, 104–120.

[2] Aizenman, M., Burchard, A., Hölder regularity and dimension bounds for random curves. *Duke Math. J.* **99** (1999), 419–453.

[3] Aizenman, M., Burchard, A., Newman, C. M., Wilson, D. B., Scaling limits for minimal and random spanning trees in two dimensions. *Random Structures Algorithms* **15** (1999), 319–367.

[4] Александров, И. А., *Параметрические продолжения в теории однолистных функций* (Aleksandrov, I. A., *Parametric continuations in the theory of univalent functions*). Izdat. "Nauka", Moscow 1976.

[5] Bauer, M., Bernard, D., 2D growth processes: SLE and Loewner chains. 2006, arXiv:math-ph/0602049.

[6] Baxter, R. J., *Exactly solved models in statistical mechanics*. Academic Press, London 1982.

[7] Beffara, V., The dimension of the SLE curves. 2002, arXiv:math.Pr/0211322.

[8] Camia, F., Newman, C. M., The full scaling limit of two-dimensional critical percolation. 2005, arXiv:math.Pr/0504036.

[9] Cardy, J. L., Critical percolation in finite geometries. *J. Phys. A* **25** (1992), L201–L206.

[10] Cardy, J., SLE for theoretical physicists. *Ann. Physics* **318** (2005), 81–118.

[11] Grimmett, G., *Percolation*. Second edition, Grundlehren Math. Wiss. 321, Springer-Verlag, Berlin 1999.

[12] Grimmett, G., *The Random-Cluster Model*. Grundlehren Math. Wiss. 333, Springer-Verlag, Berlin 2006.

[13] Kakutani, Sh., Two-dimensional Brownian motion and harmonic functions. *Proc. Imp. Acad. Tokyo* **20** (1944), 706–714.

[14] Kaufman, B., Onsager, L., Crystal Statistics. III. Short-Range Order in a Binary Ising Lattice. *Phys. Rev.* **76** (1949), 1244–1252.

[15] Kager, W., Nienhuis, B., A guide to stochastic Löwner evolution and its applications. *J. Statist. Physics*, **115** (2004), 1149–1229.

[16] Kenyon, R., Conformal invariance of domino tiling. *Ann. Probab.* **28** (2000), 759–795.

[17] Kenyon, R., Dominos and the Gaussian free field. *Ann. Probab.* **29** (2001), 1128–1137.

[18] Kesten, H., *Percolation Theory for Mathematicians*. Progr. Probab. Statist.2, Birkhäuser, Boston 1982.

[19] Kesten, H., Scaling relations for 2D-percolation. *Comm. Math. Phys.* **109** (1987), 109–156.

[20] Langlands, R., Pouliot, Ph., Saint-Aubin, Y., Conformal invariance in two-dimensional percolation. *Bull. Amer. Math. Soc.* (*N.S.*) **30** (1994), 1–61.

[21] Langlands, R. P., Lewis, M.-A., Saint-Aubin, Y., Universality and conformal invariance for the Ising model in domains with boundary. *J. Statist. Phys.* **98** (2000), 131–244.

[22] Lawler, G. F., *Conformally Invariant Processes in the Plane*. Math. Surveys Monogr. 114, Amer. Math. Soc., Providence, RI, 2005.

[23] Lawler, G. F., Schramm, O., Werner, W., Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab.* **32** (2004), 939–995.

[24] Lévy, P., *Processus Stochastiques et Mouvement Brownien. Suivi d'une Note de M. Loève.* Gauthier-Villars, Paris 1948.

[25] Löwner, K., Untersuchungen über schlichte konforme Abbildung des Einheitskreises, I. *Math. Ann.* **89** (1923), 103–121.

[26] Madras, N., Slade, G., *The Self-Avoiding Walk*. Probab. Appl., Birkhäuser, Boston 1993.

[27] McCoy, B. M., Wu, T. T., *The two-dimensional Ising model*. Harvard University Press, Cambridge, MA, 1973.

[28] Nienhuis, B., Exact critical point and critical exponents of O(n) models in two dimensions. *Phys. Rev. Lett.* **49** (1982), 1062–1065.

[29] Nienhuis, B., Coulomb gas description of 2-D critical behaviour. *J. Statist. Phys.* **34** (1984), 731–761.

[30] Rohde, S., Schramm, O., Basic properties of SLE. *Ann. of Math.* **161** (2005), 883–924.

[31] Russo, L., A note on percolation. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **43** (1978), 39–48.

[32] Schramm, O., Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118** (2000), 221–288.

[33] Schramm, O., A percolation formula. *Elect. Comm. Probab.* **6** (2001) 115–120.

[34] Schramm, O., Conformally invariant scaling limits, an overview and a collection of problems. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2006/2007.

[35] Schramm, O., Sheffield, S., Harmonic explorer and its convergence to SLE_4. *Ann. Probab.* **33** (2005), 2127–2148.

[36] Seymour, P. D., and Welsh, D. J. A., Percolation probabilities on the square lattice. *Ann. Discrete Math.* **3** (1978), 227–245.

[37] Smirnov, S., Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris* **333** (2001), 239–244.

[38] Smirnov, S., Critical percolation in the plane. I. Conformal invariance and Cardy's formula. II. Continuum scaling limit. *Preprint*, 2001.

[39] Smirnov, S., Conformal invariance of 2D Ising model. *Preprint*, 2006.

[40] Smirnov, S., Werner, W., Critical exponents for two-dimensional percolation. *Math. Res. Lett.* **8** (2001), 729–744.

[41] Werner, W., Random planar curves and Schramm-Löwner evolutions. In *Lectures on probability theory and statistics*, Lecture Notes in Math. 1840, Springer-Verlag, Berlin 2004, 107–195.

[42] Werner, W., Some recent aspects of random conformally invariant systems. 2005, arXiv: math.PR/0511268.

[43] Wiener, N., Differential space. *J. Math. Phys.* **58** (1923), 131–174.

[44] Yang, C. N., The spontaneous magnetization of a two-dimensional Ising model. *Phys. Rev.* (2) **85** (1952), 808–816.

Université de Genève, Section de Mathématiques, 2–4, rue du Lièvre, Case postale 64,
1211 Genève 4, Switzerland
E-mail: smirnov@math.unige.ch

Aspects of the L^2-Sobolev theory of the $\bar\partial$-Neumann problem

Emil J. Straube*

Abstract. The $\bar\partial$-Neumann problem is the fundamental boundary value problem in several complex variables. It features an elliptic operator coupled with non-coercive boundary conditions. The problem is globally regular on many, but not all, pseudoconvex domains.

We discuss several recent developments in the L^2-Sobolev theory of the $\bar\partial$-Neumann problem that concern compactness and global regularity.

Mathematics Subject Classification (2000). Primary 32W05; Secondary 35N15.

Keywords. $\bar\partial$-Neumann problem, regularity in Sobolev spaces, compactness, pseudoconvex domains.

1. Introduction

The $\bar\partial$-Neumann problem was formulated in the fifties by D. C. Spencer as a means to generalize the theory of harmonic integrals (i.e. Hodge theory) to non-compact complex manifolds. For domains in $\mathbb{C}^n$, which is the context we will restrict ourselves to almost exclusively in this paper, the problem can be formulated as follows. Denote by Ω a pseudoconvex domain in $\mathbb{C}^n$, and by $L^2_{(0,q)}(\Omega)$ the space of $(0,q)$-forms on Ω with square integrable coefficients. Each such form can be written uniquely as a sum

$$u = {\sum_J}' u_J d\bar z_J, \tag{1.1}$$

where $J = (j_1, \dots, j_q)$ is a multi-index with $j_1 < j_2 < \cdots < j_q$, $d\bar z_J = d\bar z_{j_1} \wedge \cdots \wedge d\bar z_{j_q}$, and the $'$ indicates summation over increasing multi-indices. The inner product

$$(u, v) = \Big({\sum_J}' u_J d\bar z_J, \, {\sum_J}' v_J d\bar z_J \Big) = {\sum_J}' \int_\Omega u_J \bar v_J \tag{1.2}$$

turns $L^2_{(0,q)}(\Omega)$ into a Hilbert space. Set

$$\bar\partial\Big({\sum_J}' u_J d\bar z_J \Big) = \sum_{j=1}^n {\sum_J}' \frac{\partial u_J}{\partial \bar z_j} d\bar z_j \wedge d\bar z_J, \tag{1.3}$$

*Supported in part by NSF grant DMS 0500842 and by a Texas A&M University Faculty Development Leave.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

where the derivatives are computed as distributions, and the domain of $\bar\partial$ is defined to consist of those $u \in L^2_{(0,q)}(\Omega)$ where the result is a $(0, q+1)$-form with square integrable coefficients. Then $\bar\partial = \bar\partial_q$ is a closed, densely defined operator from $L^2_{(0,q)}(\Omega)$ to $L^2_{(0,q+1)}(\Omega)$, and as such has a Hilbert space adjoint. This adjoint is denoted by $\bar\partial^*_q$. (We will not use the subscripts when the form level at which the operators act is clear or not an issue.) One can check that $\bar\partial\bar\partial = 0$, so that we arrive at a complex, the $\bar\partial$ (or Dolbeault)-complex:

$$L^2(\Omega) \xrightarrow{\bar\partial} L^2_{(0,1)}(\Omega) \xrightarrow{\bar\partial} L^2_{(0,2)}(\Omega) \xrightarrow{\bar\partial} \cdots \xrightarrow{\bar\partial} L^2_{(0,n)}(\Omega) \xrightarrow{\bar\partial} 0.$$

In analogy to the Laplace–Beltrami operator associated to the DeRham complex on a Riemannian manifold, one forms the complex Laplacian

$$\Box_q = \bar\partial_{q-1}\bar\partial^*_{q-1} + \bar\partial^*_q\bar\partial_q, \tag{1.4}$$

with domain so that the compositions are defined. The $\bar\partial$-complex is elliptic. The $\bar\partial$-Neumann problem is the problem of inverting $\Box_q$; that is, given $v \in L^2_{(0,q)}(\Omega)$, find $u \in \mathrm{Dom}(\Box_q)$ such that $\Box_q u = v$. Note that $\mathrm{Dom}(\Box_q)$ involves the two boundary conditions $u \in \mathrm{Dom}(\bar\partial^*)$ and $\bar\partial u \in \mathrm{Dom}(\bar\partial^*)$; these are the $\bar\partial$-Neumann boundary conditions. The condition $u \in \mathrm{Dom}(\bar\partial^*)$ is equivalent to a Dirichlet condition for the (complex) normal component of u. Similarly, the condition $\bar\partial u \in \mathrm{Dom}(\bar\partial^*)$ is equivalent to a Dirichlet condition on the normal component of $\bar\partial u$, that is, a *complex (or $\bar\partial$-) Neumann condition* for u.

From the point of view of partial differential equations, the $\bar\partial$-Neumann problem represents the prototype of a problem where the operator is elliptic, but the boundary conditions are not coercive, so that the classical elliptic theory does not apply. From the point of view of several complex variables, the importance of the problem stems from the fact that its solution provides a Hodge decomposition in the context of the $\bar\partial$-complex, together with the attendant elegant machinery (as envisioned by Spencer). For example, such a decomposition readily produces a solution to the inhomogeneous $\bar\partial$ equation, as follows. Assume for the moment that $\Box_q$ has a (bounded) inverse in $L^2_{(0,q)}(\Omega)$, say N_q. Then we have the orthogonal decomposition

$$u = \bar\partial\bar\partial^* N_q u + \bar\partial^*\bar\partial N_q u, \quad u \in L^2_{(0,q)}(\Omega). \tag{1.5}$$

If $\bar\partial u = 0$, then $\bar\partial^*\bar\partial N_q u$ is $\bar\partial$-closed as well (from (1.5)). Consequently, $\bar\partial^*\bar\partial N_q u = 0$ (since it is also orthogonal to $\mathrm{Ker}(\bar\partial)$), and

$$u = \bar\partial\big(\bar\partial^* N_q u\big), \tag{1.6}$$

with $\|\bar\partial^* N_q u\|^2 = (\bar\partial\bar\partial^* N_q u, N_q u) \le C\|u\|^2$. Thus the operator $\bar\partial^* N$ provides an L^2-bounded solution operator to $\bar\partial$. In fact, this operator gives the (unique) solution

orthogonal to $\mathrm{Ker}(\bar{\partial})$ (equivalently: the solution with minimal norm). This solution is called the canonical (or Kohn) solution.

That $\Box_q$ *does* have a bounded inverse N_q on bounded pseudoconvex domains was known by the mid sixties. Kohn ([64], [65], [66]) solved the $\bar{\partial}$-Neumann problem for strictly pseudoconvex domains, showing that in this case, not only is there an L^2-bounded inverse, but N_q exhibits a subelliptic gain of one derivative as measured in the L^2-Sobolev scale. Another interesting approach was given by Morrey in [79]. Hörmander ([58], see also Andreotti–Vesentini [1] for similar techniques) proved certain Carleman type estimates which in the case of bounded pseudoconvex domains imply the existence of N_q as a bounded self-adjoint operator on $L^2_{(0,q)}(\Omega)$. Early applications included embedding of real analytic manifolds ([78], [79]), a new solution of the Levi problem ([65]), a new proof of the Newlander–Nirenberg theorem on integrable almost complex manifolds ([65]), and in general, an approach to several complex variables which takes advantage of the then newly developed $\bar{\partial}$-methods ([58], [59]). Interesting 'eyewitness' accounts of this foundational period by two of the principals appear in [60] and [69], respectively.

It is not hard to see that Kohn's results for strictly pseudoconvex domains are optimal: N can never gain more than one derivative, and it can gain one derivative only when the domain is strictly pseudoconvex. However, under what circumstances subellipticity with a fractional gain of less than one derivative holds was not understood until the early eighties. Kohn gave sufficient conditions in [67], satisfied for example when the boundary is real-analytic [40]. In deep work, his students Catlin ([18], [19], [21]) and D'Angelo ([28], [29], [30]) resolved this question: on a smooth bounded pseudoconvex domain in $\mathbb{C}^n$, the $\bar{\partial}$-Neumann problem is subelliptic if and only if each boundary point is of finite type,that is, the order of contact, at the point, of complex varieties with the boundary is finite. For more on these ideas, see [32], [34], [33], [70].

When N_q does not gain derivatives, but is still compact (as an operator on $L^2_{(0,q)}(\Omega)$), it follows from work of Kohn and Nirenberg ([71]) that N_q preserves the Sobolev spaces $W^s_{(0,q)}(\Omega)$ for all $s \geq 0$. In particular, N_q preserves $C^\infty_{(0,q)}(\overline{\Omega})$ (it is globally regular). Work of Catlin ([20], compare also Takegoshi [95]) and Sibony ([84]) shows that compactness provides indeed a viable route to global regularity: the compactness condition can be verified on large classes of domains. We refer the reader to [47] for a survey of compactness in the $\bar{\partial}$-Neumann problem. In Section 2 below, we will discuss some developments that have occurred since the publication of [47].

In addition to the 'usual' (pde) reasons for studying regularity properties of a differential operator, there are, in the case of the $\bar{\partial}$-Neumann problem, the implications global regularity of the $\bar{\partial}$-Neumann operator has for several complex variables. Chief among these is the relevance for boundary behavior of biholomorphic or proper holomorphic maps. Namely, if Ω_1 and Ω_2 are two bounded pseudoconvex domains in $\mathbb{C}^n$ with smooth boundaries, such that the $\bar{\partial}$-Neumann operator on $(0, 1)$-forms (i.e. N_1)

on Ω_1 is globally regular, then any proper holomorphic map from Ω_1 to Ω_2 extends smoothly to the boundary of Ω_1 ([7], [41]). This is a highly nontrivial result: in contrast to the one variable situation, where the result is classical, the general case in higher dimensions, even for biholomorphic maps, is open. An exposition of the ideas and issues involved here can be found in [6], [72], [23].

That global regularity holds on large classes of pseudoconvex domains where local regularity or compactness fail was shown in the early nineties by Boas and the author ([11], [13]). They proved in [11] that if Ω admits a defining function whose complex Hessian is positive semi-definite at points of the boundary (a condition slightly more restrictive than pseudoconvexity), then the $\bar\partial$-Neumann problem is globally regular (for all q). This class of domains includes in particular all (smooth) convex domains. The proof is based on the existence of certain families of vector fields which have good approximate commutator properties with $\bar\partial$. In [13], the authors studied the situation when the boundary points of infinite type form a complex submanifold (with boundary) of the boundary of the domain. They identified a DeRham cohomology class associated to the submanifold as the obstruction to the existence of the vector fields needed. In particular, a simply connected complex manifold in the boundary is benign for global regularity of the $\bar\partial$-Neumann problem. It is noteworthy that this cohomology class also plays a role in deciding whether or not the closure of the domain admits a Stein neighborhood basis ([5]).

The question whether global regularity holds on general pseudoconvex domains turned out to be very difficult and was resolved only in the mid nineties. Barrett ([3], see also [2] and [63] for predecessors) showed that on the worm domains of Diederich and Fornæss ([38]), N_1 does not preserve $W^s_{(0,1)}(\Omega)$ for s sufficiently large, depending on the winding (that is, exact regularity fails). Christ ([24], see also [25], [26]) resolved the question by proving certain a priori estimates for N_1 on these domains that would imply exact regularity in Sobolev spaces (and thus would contradict Barrett's result) if N_1 were to preserve the space of forms smooth up to the boundary.

In Section 3, we will discuss some recent developments. In [93] and [45], the authors consider the case where the boundary is finite type except for a Levi-flat piece which is 'nicely' foliated by complex hypersurfaces. Whether or not the families of vector fields with good approximate commutator properties with $\bar\partial$ exist turns out to be equivalent to a property of the Levi foliation much studied in foliation theory, namely whether or not the foliation can be defined globally by a closed one-form. Sucheston and the author showed in [92] that the approaches via plurisubharmonic defining functions and vector fields with good approximate commutator properties with $\bar\partial$ are actually equivalent, when suitably reformulated. This left two main avenues to global regularity: compactness and plurisubharmonic defining functions and/or good vector fields. These two approaches were unified by the author in [91], via a new sufficient condition for global regularity.

Detailed accounts of the $\bar\partial$-Neumann theory, from different points of view, may be found in [43], [59], [72], [23], [73], [81]. Developments up to about ten years ago

are covered in [14] (for compactness, see [47]). Two recent informative surveys that concentrate on topics not covered here are the following: [83] deals with estimates on Lipschitz domains, and [77] discusses applications obtained by inserting a so called twisting factor into the $\bar\partial$-complex.

Although this paper concentrates on aspects of the $\bar\partial$-Neumann problem on domains in $\mathbb{C}^n$, we wish to mention a very important and fruitful development, namely the application of L^2-methods to algebraic and complex geometry, and vice versa. For expositions of this very active area of research, we refer the reader to [35], [36], [86], [87].

2. Compactness

It is of interest in several contexts to know whether or not the $\bar\partial$-Neumann operator is, or is not, compact. Examples include global regularity ([71]), the Fredholm theory of Toeplitz operators(see e.g. [54]), and existence or non-existence of Henkin–Ramirez type kernels for solving $\bar\partial$ ([51]). A fairly comprehensive discussion of compactness, up to about 1999, is in [47]. There one also finds complete proofs and/or references for background material. In this section, we review some recent developments.

2.1. Sufficient conditions for compactness. In [20] Catlin introduced a sufficient condition for compactness which he called property (P). The boundary of a domain is said to satisfy property (P) if for every positive number M there are an open neighborhood U_M of $b\Omega$ and a plurisubharmonic function $\lambda_M \in C^2(U_M \cap \Omega)$ with $0 \le \lambda_M \le 1$, such that for all $z \in U_M \cap \Omega$,

$$\sum_{j,k}^{n} \frac{\partial^2 \lambda_M}{\partial z_j \partial \overline{z}_k}(z) w_j \overline{w}_k \ge M|w|^2. \tag{2.1}$$

(This is somewhat more general than the formulation given in [20], but see also [47].) Property (P) can be very nicely reformulated, on Lipschitz domains, in the spirit of Oka's lemma: the boundary satisfies property (P) if and only if it locally admits functions ρ comparable to minus the boundary distance, such that the complex Hessian of $-\log(-\rho)$ tends to infinity upon approach to the boundary ([50]). If the boundary of the bounded pseudoconvex domain satisfies property (P), then the $\bar\partial$-Neumann operator N_q on Ω is compact, $1 \le q \le n$ ([20], [88] when no boundary regularity is assumed). There are natural versions of property (P) for $(0, q)$-forms, see [47]: (2.1) is replaced by the requirement that the sum of the smallest q eigenvalues of the complex Hessian of λ_M should be at least M. Note that then $P = P_1 \Rightarrow P_2 \Rightarrow \cdots \Rightarrow P_n$. This is appropriate, since compactness of N_q likewise percolates up the complex: if N_q is compact, then so is N_{q+1} (an observation due to McNeal ([76]), see also the proof of Lemma 2 in [91]). A detailed study of property (P_1) is in [84],where various equivalent characterizations are given (with some minimal boundary regularity required for

equivalence to the definition we have adopted here, see [47], section 3). In particular, $b\Omega$ satisfies property (P_1) if and only if every continuous function on $b\Omega$ can be approximated uniformly on $b\Omega$ by functions plurisubharmonic in a neighborhood of $b\Omega$.

In $\bar\partial$-problems, the condition that a function be bounded can sometimes be replaced by the condition that the gradient of the function be bounded in the metric induced by the complex Hessian of the function (compare [9] and the references there). In the present context, this was realized by McNeal ([75]). Say that the boundary of the domain Ω satisfies condition $(\tilde P_q)$ if, for every $M > 0$, there exists $\lambda_M \in C^2(U_M \cap \Omega)$, where U_M is an open neighborhood of $b\Omega$, such that the sum of any q eigenvalues of its complex Hessian is at least M, and

$$\sum_K{}' \Big| \sum_{j=1}^n \frac{\partial \lambda_M}{\partial z_j}(z) w_{jK} \Big|^2 \le C \sum_K{}' \sum_{j,k=1}^n \frac{\partial^2 \lambda_M}{\partial z_j \partial \bar z_k}(z) w_{jK} \bar w_{kK} \tag{2.2}$$

for all $w \in \Lambda_z^{(0,q)}$ and $z \in U_M \cap \Omega$, where $\Lambda_z^{(0,q)}$ denotes the space of $(0,q)$-forms at z. (Actually, for $q > 1$, this definition is slightly more general than McNeal's; in particular, it does not force λ_M to be plurisubharmonic.) McNeal ([75]) proved the following theorem.

Theorem 2.1. *Let Ω be a bounded pseudoconvex domain in $\mathbb{C}^n$, $1 \le q \le n$. If Ω satisfies condition $(\tilde P_q)$, then N_q is compact.*

Again note that $\tilde P_1 \Rightarrow \tilde P_2 \Rightarrow \cdots \Rightarrow \tilde P_n$ (compare the proof of Lemma 2 in [91].) Also, there is a weak formulation of (2.2) in terms of currents which is sufficient for Theorem 2.1, see [75] for details.

To prove Theorem 2.1, one uses that compactness of N_q is equivalent to compactness of the canonical solution operators $\bar\partial^* N_q$ and $\bar\partial^* N_{q+1}$ (or their adjoints, see e.g. [47], Lemma 1.1). Since $(\tilde P_q) \Rightarrow (\tilde P_{q+1})$, it suffices to establish compactness of $\bar\partial^* N_q$. We sketch an argument (slightly different from McNeal's) to show how the hypotheses of the theorem enter (as well as for use in Remark 2.2 below). We only consider $(0,1)$-forms for simplicity. Also, we assume that Ω is smooth and that λ_M is smooth up to the boundary (see [88] and [47] for the regularization procedure in the non-smooth case). We may extend λ_M smoothly to all of Ω (by shrinking U_M, where the estimates hold). The starting point is the classical Morrey–Kohn–Hörmander inequality ([23], Proposition 4.3.1, and the density Lemma 4.3.2). Taking the weight function to be λ_M and computing the adjoint of $\bar\partial^*$ in the weighted metric in terms of $\bar\partial^*$ and terms of order zero in u gives for $u \in \mathrm{Ker}(\bar\partial) \cap \mathrm{Dom}(\bar\partial^*)$

$$\int_\Omega \sum_{j,k} \frac{\partial^2 \lambda_M}{\partial z_j \partial \bar z_k} u_j \bar u_k e^{-\lambda_M} \lesssim \|\bar\partial^*(e^{-\lambda_M/2} u)\|^2 + \Big\| e^{-\lambda_M/2} \sum_j \frac{\partial \lambda_M}{\partial z_j} u_j \Big\|^2. \tag{2.3}$$

The constant in (2.2) may be assumed as small as we wish (by scaling $\lambda \to t\lambda$, see [75], p. 199). Therefore, the last term in (2.3) can be absorbed into the left hand side,

for $z \in U_M$. The result is, upon also applying (2.1), using interior elliptic regularity to estimate $\|e^{-\lambda_M/2}u\|^2_{\Omega\setminus\bar U_M}$ (this term controls the various error terms), and absorbing terms:

$$\|e^{-\lambda_M/2}u\|^2 \lesssim \frac{1}{M}\|\bar\partial^*(e^{-\lambda_M/2}u)\|^2 + C_M\|e^{-\lambda_M/2}u\|^2_{-1}, \tag{2.4}$$

when $u \in \mathrm{Ker}(\bar\partial) \cap \mathrm{Dom}(\bar\partial^*)$. Denote by B the standard Bergman projection, and by B_φ the Bergman projection in the weighted space with weight e^φ. Let $v \in \mathrm{Ker}(\bar\partial)$. Then $v = B(e^{-\lambda_M/2}u)$, with $u = B_{-\lambda_M/2}(e^{\lambda_M/2}v) \in \mathrm{Ker}(\bar\partial)$. We obtain

$$\|v\|^2 \lesssim \frac{1}{M}\|\bar\partial^* v\|^2 + C_M\|e^{-\lambda_M/2}B_{-\lambda_M/2}(e^{\lambda_M/2}v)\|^2_{-1}. \tag{2.5}$$

For M fixed, the norm in the last term on the right hand side of (2.5) is compact with respect to $\|v\|$, hence with respect to $\|\bar\partial^* v\|$ (the canonical solution operator to $\bar\partial^*$ is continuous in L^2). Since M is arbitrary, this implies that $(\bar\partial^* N_1)^*$, the canonical solution operator to $\bar\partial^*$, is compact (see e.g [22], Lemma 1, [75], Lemma 2.1). This concludes the (sketch of) proof of Theorem 2.1.

It is easy to see that P_q implies $\tilde P_q$, by considering the family $\mu_M := e^{\lambda_M}$, $M > 0$ (see [75] for details). The exact relationship is however not understood. But there are two classes of domains where property (P_q) is known to be equivalent to compactness of N_q, and hence condition $(\tilde P_q)$ is also equivalent to both compactness of N_q and property (P_q). These are the bounded locally convexifiable domains (with no boundary smoothness assumptions assumed beyond what is implied by local convexifiability (Lipschitz), [46], [47]) and the smooth bounded Hartogs domains in $\mathbb{C}^2$ ([27]). In addition, (P) and $(\tilde P)$ are known to agree for planar domains ([48], Lemma 7). Although condition $(\tilde P_1)$ also appears naturally in connection with another important question in several complex variables, namely with that of the existence of a Stein neighborhood basis for the closure of the domain ([84], (4.4) on p. 317), it is not well studied at all. This is in stark contrast to property (P_q), where we have Sibony's theory ([84], see also [47] when $q > 1$). It would be very interesting to have an analogous treatment of condition $(\tilde P_q)$. Finally, its connection with Stein neighborhoods suggests studying what the implications of compactness in the $\bar\partial$-Neumann problem are for the existence of a Stein neighborhood basis for the closure.

Compactness can be viewed as a limiting case of subellipticity. Subellipticity is equivalent to having a bounded plurisubharmonic function, near the boundary, whose Hessian blows up like a power of the reciprocal of the boundary distance ([21], [88], compare also [50], [55]). The only way to show the direction *subellipticity* $\Rightarrow$ *good plurisubharmonic functions* that the author is aware of is via finite type of the boundary. It would be very interesting to have a direct proof of this fact, arguing directly from the subelliptic estimates and bypassing the geometric arguments related to finite type. If there is a reasonable characterization of compactness in terms of (potential theoretic) properties of the boundary (such as $P/\widetilde{P}$ or a slightly weaker

property), it should emerge from such a proof when one extracts the features that remain valid in the limiting case.

Remark 2.2. Takegoshi gives a sufficient condition for compactness in [95] which is a precursor to condition $(\tilde{P}_1)$, in the sense that it replaces the uniform boundedness condition on the functions in property (P_1) by a boundedness condition on the gradients. In fact, Takegoshi's condition implies $(\tilde{P}_1)$, but with (2.2) only for complex tangential w's. But this is enough to prove Theorem 2.1, because the forms to which one needs to apply the estimates are in the domain of $\bar{\partial}^*$ (see (2.3) above), so they are complex tangential (at points of the boundary, but the normal component of a form satisfies a subelliptic estimate, and so terms caused by it are under control).

We next want to describe a technique for establishing compactness that does not rely on (P)/$(\tilde{P})$, introduced recently by the author ([90]). Such a technique is of interest because it is not understood how much stronger (if at all) $(\tilde{P})$ is than compactness. Also, as will be seen, for domains in $\mathbb{C}^2$, this technique yields a sufficient condition that modulo a certain (albeit crucial) lower bound on certain radii is also necessary.

Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^2$. If Z is a (real) vector field defined in some open subset of $b\Omega$ (or of $\mathbb{C}^2$), we denote by $\mathcal{F}_Z^t$ the flow generated by Z. In $\mathbb{C}^2$, the various notions of finite type coincide (see [32]), so we do not need to specify which notion we mean. Recall that the set of infinite type points in the boundary is compact. Finally, $B(P, r)$ denotes the open ball of radius r centered at P. The following theorem is the main result in [90].

Theorem 2.3. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^2$. Denote by K the set of boundary points of infinite type. Assume there exist constants C_1, $C_2 > 0$, and C_3 with $1 \leq C_3 < 3/2$, and a sequence $\{\varepsilon_j > 0\}_{j=1}^{\infty}$ with $\lim_{j\to\infty} \varepsilon_j = 0$ so that the following holds. For every $j \in \mathbb{N}$ and $P \in K$ there is a (real) complex tangential vector field $Z_{P,j}$ of unit length defined in some neighborhood of P in $b\Omega$ with $\max |\operatorname{div} Z_{P,j}| \leq C_1$ such that $\mathcal{F}_{Z_{P,j}}(B(P, C_2(\varepsilon)^{C_3}) \cap K) \subseteq b\Omega \setminus K$. Then the $\bar{\partial}$-Neumann operator on Ω is compact.*

The assumptions in the theorem quantify the notion that at points of K, there should exist a (real) complex tangential direction transversal to K in which $b\Omega \setminus K$ (the good set) is 'thick' enough. A geometrically very simple special case occurs when $b\Omega \setminus K$ satisfies a complex tangential cone condition (that is, the axis of the cone at a point P lies in a complex tangential direction). In this case, the assumptions in the theorem are satisfied with $C_3 = 1$ (see [90] for details).

Corollary 2.4. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^2$. Denote by K the set of boundary points of infinite type. Assume that $b\Omega \setminus K$ satisfies a complex tangential cone condition (as an open subset of $b\Omega$). Then the $\bar{\partial}$-Neumann operator is compact.*

The main idea of the proof of Theorem 2.3 is very simple. In order to derive a compactness estimate, what needs to be estimated is the L^2-norm of a form u near K.

To do this near a point P in K, we express u near P in terms of u in a patch which meets the boundary in a relatively compact subset of $b\Omega \setminus K$, plus the integral of the derivative of u in the direction $Z_{P,j}$. The first contribution is easily handled by subelliptic estimates, while the second is dominated by the length of the curve (which is ε_j) times the L^2-norm of $Z_{P,j}u$ on a certain subset of the boundary. But in $\mathbb{C}^2$, this norm is dominated by $\|\bar{\partial}u\| + \|\bar{\partial}^* u\|$, because $Z_{P,j}$ is complex tangential (so called maximal estimates hold, c.f. [37]). When one sums up over the various patches (for a fixed j), overlap as well as divergence issues arise. These are handled by the uniformity built into the assumptions.

The conditions in Theorem 2.3 are natural, and, in fact, modulo the size of the lower bound $C_2(\varepsilon_j)^{C_3}$ on the radius of the balls, necessary. Indeed, if N_1 is compact, the boundary contains no analytic discs (since we are in $\mathbb{C}^2$, see [47] for a proof). This implies, by a result of Catlin ([16], Proposition 3.1.12, see also [82], Lemma 3) that for each point $P \in K$ and for every $\varepsilon > 0$, there is a complex tangential vectorfield $Z_{P,\varepsilon}$ (of unit length) near P so that on the integral curve of $Z_{P,\varepsilon}$ through P there is a strictly pseudoconvex point at distance (measured along the curve) less than ε. Then there is a ball $B(P, r)$ which is transported, for $t = \varepsilon$, by the flow generated by $Z_{P,\varepsilon}$, into the points of finite type: it suffices to take r small enough. Since there are smooth bounded pseudoconvex domains in $\mathbb{C}^2$ without discs in their boundaries, but whose $\bar{\partial}$-Neumann operator is not compact ([74], [47]), this discussion also shows that without a lower bound on r, the conclusion of the theorem does not hold. The lower bound given in Theorem 2.3 is probably not optimal. An 'optimal' bound (if one exists), in a sense to be made precise, would be of great interest: in light of the above discussion, such a bound essentially amounts to a characterization of compactness in the $\bar{\partial}$-Neumann problem on domains in $\mathbb{C}^2$.

Theorem 2.3 does not hold in dimension $n > 2$. Consider a convex domain with a disc in its boundary. When $n > 2$, there is an additional complex tangential direction in which to flow, so that the assumptions in Theorem 2.3 can be satisfied. Yet such domains have noncompact $\bar{\partial}$-Neumann operator ([46]). Since the only place where the proof uses that the domain is in $\mathbb{C}^2$ is the invocation of maximal estimates, an obvious generalization to $\mathbb{C}^n$ is to require the domain to satisfy maximal estimates (equivalently: all the eigenvalues of the Levi form are comparable, see [37]).There is, however, a more interesting generalization in [80]. It suffices to be able to flow into the set of finite type points along curves whose tangents lie in a complex tangential direction associated with the smallest eigenvalue of the Levi form.

The author does not know examples of domains that satisfy the assumptions in Theorem 2.3, but do not satisfy condition $(\tilde{P})$. As far as just asserting compactness of the $\bar{\partial}$-Neumann problem on the domains in the theorem is concerned, it does not matter whether or not these domains always satisfy $(\tilde{P})$: we have, in any case, a simple geometric proof of compactness for these domains. However, from the point of view of understanding to what extent $(\tilde{P})$ is necessary for compactness, this question is obviously very important.

2.2. Obstructions to compactness. An analytic disc in the boundary constitutes the most blatant violation of condition $(\tilde{P})$ (on domains whose boundaries are locally the graph of a continuous function) as well as of the condition in Theorem 2.3. This is obvious for the condition in Theorem 2.3 (recall that the setup is in $\mathbb{C}^2$). For condition $(\tilde{P})$, this can be seen by pulling back (suitable translates of) the plurisubharmonic functions to the unit disc D in the plane: there do not exist subharmonic functions in D satisfying (2.1) and (2.2) for arbitrarily large M (compare Appendix A in [47]). Indeed, integration by parts and (2.2) give for $u \in C_0^\infty(D)$

$$\begin{aligned}\int_D \frac{\partial^2\lambda}{\partial z\partial\bar z}|u|^2 = \int_D \frac{\partial^2\lambda}{\partial z\partial\bar z}u\bar u &\le \left|\int_D \frac{\partial\lambda}{\partial\bar z}\frac{\partial u}{\partial z}\bar u\right| + \left|\int_D \frac{\partial\lambda}{\partial\bar z}u\frac{\partial\bar u}{\partial z}\right| \\ &\le \int_D \left|\frac{\partial\lambda}{\partial\bar z}\right|^2|u|^2 + \int_D\left|\frac{\partial u}{\partial z}\right|^2 \le C\int_D \frac{\partial^2\lambda}{\partial z\partial\bar z}|u|^2 + \int_D\left|\frac{\partial u}{\partial z}\right|^2,\end{aligned} \tag{2.6}$$

where C is the constant from (2.2). We have used here that $\|\partial u/\partial\bar z\|^2 = \|\partial u/\partial z\|^2$. As pointed out earlier, C may be taken as small as we wish. Taking a family with $C = 1/4$ in (2.2) (hence in (2.6)) and combining with (2.1) gives

$$\frac{M}{4} \le \inf_{u\in C_0^\infty(D)} \frac{\int_D \left|\frac{\partial u}{\partial z}\right|^2}{\int_D |u|^2}. \tag{2.7}$$

(The infimum on the right hand side of (2.7) is (up to a factor $1/4$) the smallest eigenvalue of the Dirichlet realization of $-\Delta$ on D.) In the case of property (P), this is of course also an obvious consequence of its characterization (see above) by the approximation property by continuous functions. A disc in the boundary is also known to be an obstruction to hypoellipticity of $\bar\partial$ ([17], [42]). It is therefore very natural to ask whether such a disc is an obstruction to compactness of the $\bar\partial$-Neumann operator.

An old folklore result, usually attributed to Catlin, says that this is indeed the case for sufficiently regular domains in $\mathbb{C}^2$. A proof for the case of Lipschitz boundary may be found in [47]. There, a simple example (the unit ball in $\mathbb{C}^2$ minus the variety $\{z_1 = 0\}$) is given that shows that some boundary regularity is needed. Whether for domains in $\mathbb{C}^2$ there can be other obstructions to compactness was resolved only surprisingly recently. Matheos ([74]) showed that there are indeed more subtle obstructions:

Theorem 2.5. *Let K be a compact subset of the complex plane with non-empty fine interior, but empty Euclidean interior. There exists a smooth bounded complete pseudoconvex Hartogs domain in $\mathbb{C}^2$ with the following properties:* (i) *its set of weakly pseudoconvex boundary points projects onto K;* (ii) *it contains no analytic discs in its boundary;* (iii) *its $\bar\partial$-Neumann operator is not compact.*

For properties of the fine topology, see e.g. [53], [48], section 3; in particular, there do exist (many) sets K as in the theorem (an explicit construction of such sets may also

be found in section 4 of [27]). The version of the theorem given here comes from [47], to where we refer the reader for details (compare also [27]). Note in particular that this also means that there are more subtle obstructions to property (P)/condition $(\tilde P)$ than discs in the boundary. This was known before, see [84].

Remark 2.6. It is easy to see that on a smooth bounded complete pseudoconvex Hartogs domain in $\mathbb{C}^2$, there is no disc in the boundary if and only if the projection of the weakly pseudoconvex boundary points has empty Euclidean interior. It will be seen in Subsection 2.3 below that the $\bar\partial$-Neumann operator is compact if and only if this set has empty fine interior (as a compact subset of $\mathbb{C}$); see the discussion following Theorem 2.11.

Remark 2.7. Whether on a smooth bounded pseudoconvex domain in $\mathbb{C}^2$ the absence of discs from the boundary implies global regularity is open.

It is folklore that the methods that work in $\mathbb{C}^2$ can be used in $\mathbb{C}^n$ to show that when the $\bar\partial$-Neumann operator N_1 is compact, the boundary cannot contain an $(n-1)$-dimensional complex manifold. However, whether a disc is necessarily an obstruction is open in general, and is arguably the most important problem concerning compactness. Şahutoğlu and the author recently showed that when the disc contains a point at which the boundary is strictly pseudoconvex in the directions transverse to the disc ([82]), then compactness does fail. This holds more generally for complex submanifolds of the boundary of arbitrary (positive) dimension.

Theorem 2.8. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^n$, $n \geq 2$. Let $p \in b\Omega$ and assume that the Levi form of $b\Omega$ at P has the eigenvalue zero with multiplicity at most k, $1 \leq k \leq n-1$ (i.e. the rank is at least $n-1-k$). If the $\bar\partial$-Neumann operator on $(0,1)$-forms is compact, then $b\Omega$ does not contain a k-dimensional complex manifold through P.*

It follows immediately from the theorem that if the set K of weakly pseudoconvex boundary points has nonempty relative interior in the boundary, then the $\bar\partial$-Neumann operator (on $(0,1)$-forms) is not compact ([82]). It suffices to observe that near a relative interior point of K where the Levi form attains its maximal rank (among points of the relative interior of K), say m, the rank has to be constant, so that the boundary is foliated there by complex manifolds of dimension $n-1-m$. Note that in general, K is considerably bigger than the set of Levi *flat* points.

The proof of Theorem 2.8 results from the following ideas. Compactness is a local property (see [47], Lemma 1.2). Therefore, it suffices to argue locally. Assume the boundary contains a complex manifold, say M. There is a holomorphic change of coordinates near P so that in the new coordinates M is affine, and the real normal to the boundary of Ω is constant on M. This is always possible ([82], Lemma 1). Next, consider a section Ω_1 of Ω through P, perpendicular to M. If Ω_1 has a subdomain Ω_2, whose boundary shares P with $b\Omega$ and such that (i) the restriction operator from the Bergman space of Ω_1 to the Bergman space of Ω_2 is not compact, and (ii) the product

$\Omega_2 \times M$ is contained in Ω (near P), then the arguments from [46] (which in turn are based on ideas from [17], [42]) carry over to produce a contradiction to the existence of a compact solution operator to $\bar\partial$ (which would be a consequence of compactness of N_1). In the situation of Theorem 2.8, any smooth subdomain Ω_2 will do, because Ω_1 is strictly pseudoconvex at P ([82], Lemma 2). If we take for Ω_2 a ball with small radius and tangent to $b\Omega_1$ at P, (ii) also holds (because the real normal to $b\Omega$ is constant along M).

Experience indicates that a flatter boundary should be even more favorable to noncompactness of the $\bar\partial$-Neumann operator. In other words, the extra assumption that the boundary is strictly pseudoconvex in the directions transverse to M should not be needed. However, the present methods do not seem to yield this. An interesting recent contribution to this circle of ideas, involving the Kobayashi metric of the domain, is in [62].

The above proof of Theorem 2.8 also raises a question of independent interest. Namely given a domain Ω and a subdomain Ω_1, when is the restriction operator from the Bergman space of Ω to that of Ω_1 compact? Of course, this happens when Ω_1 is relatively compact in Ω, so the case of interest is that where the domains share a boundary point. As mentioned above, this restriction is known not to be compact when Ω is smooth near a strictly pseudoconvex boundary point P and $b\Omega_1$ shares P with $b\Omega$ and is smooth there ([82], Lemma 2). In addition, this restriction is known not to be compact when Ω is convex, $P = 0 \in b\Omega$, and $\Omega_1 = r\Omega$, for $r < 1$ ([46]). The general situation is not understood.

2.3. Hartogs domains in $\mathbb{C}^2$ and semi-classical analysis of Schrödinger operators. It is well known that $\bar\partial$ and related operators on Hartogs domains can be studied by means of weighted operators on the base domain. In the sequel, the base domain U will be a planar domain (i.e. the Hartogs domain is in $\mathbb{C}^2$). The resulting weighted problems lead to Schrödinger operators on U, see for example [8] and the references there.

Let U be a bounded domain in $\mathbb{C}$, $\phi(z) \in C^2(\overline{U})$. Denote by S_ϕ the Schrödinger operator with magnetic potential $A = -(\partial\phi/\partial y)dx + (\partial\phi/\partial x)dy$, magnetic field $dA = \Delta\phi(dx \wedge dy)$, and electric potential $V = \Delta\phi$. That is, S_ϕ is given by (the Dirichlet realization of)

$$S_\phi = -\left[(\partial/\partial x + i\partial\phi/\partial y)^2 + (\partial/\partial y - i\partial\phi/\partial x)^2\right] + \Delta\phi. \tag{2.8}$$

Denote by S_ϕ^0 the corresponding nonmagnetic Schrödinger operator, given by (the Dirichlet realization of)

$$S_\phi^0 = -\Delta + \Delta\phi. \tag{2.9}$$

For (very) brief introductions to Schrödinger operators, we refer the reader to [48], section 2 or [27], section 2.3. For a detailed treatment in the context of semi-classical analysis relevant here, see [52].

Let Ω be a bounded complete pseudoconvex Hartogs domain in $\mathbb{C}^2$ given by $\Omega = \{(z, w) \in \mathbb{C}^2 : z \in U, |w| < e^{-\phi(z)}\}$, where U is a domain in $\mathbb{C}$. Note that

pseudoconvexity forces ϕ to be plurisubharmonic. Also, smoothness of Ω means that ϕ is smooth on U, but not on $\overline{U}$, but the notions needed here are still well defined, compare [48], [27]. The rotation invariance in the w variable brings a discrete Fourier variable into play, and so what one actually has when analyzing the $\bar{\partial}$ and related problems on Hartogs domains are sequences of Schrödinger operators of the form $\{S_{n\phi}\}_{n=1}^{\infty}$ and $\{S^0_{n\phi}\}_{n=1}^{\infty}$, respectively (see [48] for details). Compactness of the $\bar{\partial}$-Neumann operator on Ω is closely linked to the behavior of the sequence of lowest eigenvalues $\{\lambda_{n\phi}\}_{n=1}^{\infty}$ (the ground state energies) of the magnetic Schrödinger operators, while property (P) of $b\Omega$ is similarly linked to the behavior of the sequence $\{\lambda^0_{n\phi}\}_{n=1}^{\infty}$ of lowest eigenvalues of their nonmagnetic counterparts. The former idea originates with [74], the latter with [48]. The precise relationships are given in the following theorem ([48]).

Theorem 2.9. *Let $\Omega = \{(z, w) \in \mathbb{C}^2 : |w| < e^{-\phi(z)}, z \in U\}$ be a smooth bounded complete pseudoconvex Hartogs domain. Suppose that $b\Omega$ is strictly pseudoconvex on $b\Omega \cap \{w = 0\}$. Then*

(1) *$b\Omega$ satisfies property (P) if and only if $\lambda^0_{n\phi} \to \infty$ as $n \to \infty$.*

(2) *The $\bar{\partial}$-Neumann operator on Ω is compact if and only if $\lambda_{n\phi} \to \infty$.*

We remark that for the domains in Theorem 2.9, property (P) and property $(\tilde{P})$ are equivalent ([48], Lemma 6). It was already noted in [48] that for some of the implications, regularity of the boundary is not needed. For a version of Theorem 2.9 that assumes very little regularity of ϕ, see [27].

It is the limit in part (2) of Theorem 2.9 that gives rise to the terminology used in the title of this subsection. Note that $S_{n\phi} = -n^2[((1/n)(\partial/\partial x) + i(\partial\phi/\partial y))^2 + ((1/n)(\partial/\partial y) - i(\partial\phi/\partial x))^2] + n\Delta\phi$. Understanding the behavior of the ground state energy as n tends to infinity is thus analogous to understanding (modulo the factor n^2) what happens when 'Planck's constant' $h = 1/n$ tends to zero. This situation is typically referred to as semi-classical analysis in the mathematical physics literature. Mathematical physics also has its own version of (1) $\Rightarrow$ (2): Simon's diamagnetic inequality asserts that $\lambda^0_{n\phi} \leq \lambda_{n\phi}$ ([85], see also [61]). Reverse relationships, when there is some kind of domination of the magnetic eigenvalues by the nonmagnetic ones, obviously of interest in our context, are known in the physics literature as paramagnetism. For more thorough discussions of these topics, we refer again to [52], [48], and [27], and their references.

This point of view has allowed to clarify the relationship between property (P) and compactness of the $\bar{\partial}$-Neumann operator on the (special) class of Hartogs domains in $\mathbb{C}^2$. Namely, Christ and Fu recently established the paramagnetic property required for the implication (2) $\Rightarrow$ (1) in Theorem 2.9.

Theorem 2.10. *Let ϕ be subharmonic on the domain $U \subseteq \mathbb{C}$, and let $\Delta\phi$ be Hölder continuous of some positive order. If $\sup_n \lambda^0_{n\phi} < \infty$ then $\liminf_{n\to\infty} \lambda_{n\phi} < \infty$.*

Since the sequence $\{\lambda^0_{n\phi}\}_{n=1}^{\infty}$ is increasing (this is obvious from (2.9); ϕ is subharmonic), we immediately get the corollary that on the domains from Theorem 2.9, property (P) and compactness of the $\bar\partial$-Neumann operator are equivalent. In fact, combining this with some additional work, Christ and Fu ([27]) were able to handle general (not necessarily complete) Hartogs domains, thus establishing the following equivalence.

Theorem 2.11. *Let $\Omega \subseteq \mathbb{C}^2$ be a smooth bounded pseudoconvex Hartogs domain. The $\bar\partial$-Neumann operator on Ω is compact if and only if $b\Omega$ satisfies property (P).*

While we have not made an effort to state Theorem 2.11 with optimal boundary smoothness assumptions (see [27]), we point out that in view of the example mentioned before the statement of Theorem 2.5, some boundary regularity is needed for the equivalence in Theorem 2.11 to hold. If the Hartogs domain is complete, then the two properties in Theorem 2.11 are also equivalent to the set K (as defined above) having empty fine interior, by work of Sibony. Namely, $b\Omega$ satisfies property (P) if and only if K does (as a subset of $\mathbb{C}$, [84], p. 310). In turn, K satisfies property (P) if and only if it has empty fine interior ([84], Proposition 1.11).

3. Global regularity

The $\bar\partial$-Neumann operator N_q is said to be globally regular if it maps $C^\infty(0,q)(\overline{\Omega})$ (necessarily continuously) into itself. It is said to be exactly regular if it maps $W^s_{(0,q)}(\Omega)$ into itself for $s \geq 0$. Exact regularity implies of course global regularity. So far, in all instances where one can prove global regularity, one actually proves exact regularity. On the worm domains, failure of global regularity ([24]) is proved via failure of exact regularity ([3]): for most s, exact a priori estimates hold in $W^s(\Omega)$, and global regularity would then give exact regularity. It is consistent with what is known that such a priori estimates might hold on all domains (they are also known to hold on the nonpseudoconvex counterexample domains from [2], see [12]).

For a survey of results up to about ten years ago, we refer he reader to [14]. In this section, we first discuss regularity on domains whose boundary contains an open patch foliated by complex hypersurfaces ([93], [45]. In Subsection 3.2, we describe a unified approach to global regularity ([92], [91]).

We recall the following important 1-form on the boundary of a domain. Let Ω be a smooth bounded pseudoconvex domain. Denote by η a purely imaginary nowhere vanishing 1-form on $b\Omega$ that annihilates the complex tangent space and its conjugate. Let T denote the purely imaginary vector field on $b\Omega$ orthogonal to the complex tangent space and its conjugate and such that $\eta(T) \equiv 1$. The real 1-form α is defined by $\alpha = -\mathcal{L}_T\eta$, the Lie derivative of η in the direction of T (compare [31], [32]). The form arises naturally in the computation of (normal components of) commutators of vector fields; indeed, if $\eta = \partial\rho - \bar\partial\rho$, and $\overline{X}$ is a local section of $T^{0,1}(b\Omega)$,

then $\alpha(\overline{X}) = 2\partial\rho([L_n, \overline{X}])$ (L_n is the complex normal). The cohomology class on complex submanifolds of the boundary mentioned in the introduction in connection with [13] is the class of α.

3.1. A foliation in the boundary. Background on foliation theory and notions used here can be found in [15] and [96], as well as in [93], [45], and their references. Assume now there is a codimension one foliation in the boundary, say the relative interior of the set K of weakly pseudoconvex points is foliated by complex manifolds of dimension $n-1$. Note that such a foliation is always transversely orientable (by the vector field T defined on all of $b\Omega$). In order to run the machinery from [11], [13], one needs a function h smooth in a relative neighborhood of K, satisfying

$$dh|_L = \alpha|_L, \quad \text{for all leaves } L. \tag{3.1}$$

For details, see [93]. Of course, this requires that the restriction of α to a leaf is closed. This does indeed hold: $d\alpha|_{\mathcal{N}_P} = 0$ always, where $\mathcal{N}_P$ is the null space of the Levi form at P, see the lemma in section 2 of [13]. Thus solving (3.1) is always possible locally. Globally, topological constraints arise. Also, the boundary behavior of h on K needs to be controlled.

It turns out that these issues are very much related to ones studied in foliation theory. Note that the foliation can be defined by η: the tangent planes to the leaves are given by the null space of η. Then the Frobenius condition reads $d\eta \wedge \eta = 0$. Hence $d\eta = \beta \wedge \eta$ for some 1-form β. α is such a form, that is

$$d\eta = \alpha \wedge \eta \quad \text{on } K \tag{3.2}$$

([96], Proposition 2.2). With (3.2), solving (3.1) is easily tied to an important property in foliation theory ([93]).

Lemma 3.1. (3.1) *can be solved* (*say on the relative interior of* K) *if and only if the Levi foliation of* K *can be defined* globally *by a* closed 1*-form.*

Indeed, if ω is a 1-form defining the foliation, then $\omega = e^{-h}\eta$, and

$$d\omega = d(e^{-h}\eta) = e^{-h}(-dh \wedge \eta + d\eta) = e^{-h}(-dh + \alpha) \wedge \eta. \tag{3.3}$$

Therefore,

$$d\omega = 0 \Leftrightarrow (-dh + \alpha) \wedge \eta = 0 \Leftrightarrow -dh|_L + \alpha|_L = 0. \tag{3.4}$$

In addition to closedness of $\alpha|_L$, solvability of (3.1) also requires that the De Rham cohomology class of $\alpha|_L$ vanishes. This again fits nicely into the foliation framework: this cohomology class coincides with the infinitesimal holonomy of L ([93], Remark 2, [15], Example 2.3.15).

We first present a result in $\mathbb{C}^2$ from [93]. The relative boundary of K in $b\Omega$, say Γ, is assumed smooth, and so is a smooth compact orientable surface embedded in $\mathbb{C}^2$.

Recall that a complex tangency at a point of Γ is called generic if it is either elliptic or hyperbolic (see [93] for more information). Note that at a hyperbolic point there are locally two leaves that meet.

Theorem 3.2. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^2$. Suppose that the set K of infinite type points of $b\Omega$ is smoothly bounded (in $b\Omega$) and that its boundary Γ is connected and has only isolated generic complex tangencies. Assume that the two leaves meeting at a hyperbolic point are distinct globally and that they have no other hyperbolic points in their closure (in K). If each leaf of the Levi foliation is closed (in the relative interior of K) and has trivial infinitesimal holonomy, then the $\bar\partial$-Neumann operator on Ω is continuous on $W^s_{(0,1)}(\Omega)$ for $s \geq 0$.*

If one assumes that the Levi foliation of K is part of a foliation of a bigger smooth Levi flat hypersurface M, with $M \cap \overline{\Omega} = K$, then boundary behavior of the leaves is easier to control, and one needs no conditions on the boundary of K. This results in a geometrically appealing sufficient condition. The following theorem is from [45]. A codimension one foliation is called simple if through every point there exists a local transversal (a line) that meets each leaf at most once.

Theorem 3.3. *Let $\Omega \subseteq \mathbb{C}^n$ be a smooth bounded pseudoconvex domain such that the set K of all boundary points of infinite D'Angelo-type is the closure of its relative interior in $b\Omega$. Assume K is contained in a smooth Levi-flat (open) hypersurface $M \subset \mathbb{C}^n$, whose Levi foliation satisfies one (hence both) of the following equivalent conditions:* (i) *the leaves of the restriction of the foliation to a neighborhood of K are topologically closed;* (ii) *the foliation is simple in a neighborhood of K. Then the $\bar\partial$-Neumann operators N_q on Ω are continuous in $W^s_{(0,q)}(\Omega)$ for $s \geq 0$, $1 \leq q \leq n$.*

Very roughly speaking, in both Theorems 3.2 and 3.3, one would like to obtain a closed form that defines the foliation by pulling back from the leaf space (which is one dimensional) a form roughly like dx. One then has to deal with the non-Hausdorff nature of this space. In Theorem 3.2, one also has to control the boundary behavior.

It is interesting to note that the main concern in [45] is not the $\bar\partial$-Neumann problem, but rather holomorphic convexity properties of compact subsets of M. The authors use asymptotically pluriharmonic defining functions for M (near a compact subset) for constructing Stein neighborhoods, and whether such defining functions exist leads precisely to the question whether the foliation, near the compact set (globally), can be defined by a closed 1-form. In view of Lemma 3.1, this is related to the equivalence between 'pluriharmonic defining functions' and 'exactness of α' in [92] (see Theorem 3.4 below). In a local context, compare also [4].

The two equivalent conditions in Lemma 3.1 are equivalent to a third one, given in terms of the flow generated by T ([93], Proposition 2). This is at least potentially of interest because the condition is in terms of T, (rather than the Levi foliation), which is well defined on the boundary of any smooth domain. Furthermore, this leads to a homological necessary and sufficient condition for the existence of a function

$h \in C^\infty(K)$ as above ([93], Theorem 3), in terms of foliation currents ([94], [15]) associated to T.

3.2. Sufficient conditions for global regularity. In [13], Remark 3 in section 4, the authors point out that the families of vector fields with good approximate commutator conditions with $\bar\partial$, required in their approach to global regularity ([11], [13]) can exist in situations where the domain does not admit (even a local) plurisubharmonic defining function. On the other hand, they had noted in [11] that it suffices to have the commutator conditions with components of $\bar\partial$ in directions that lie in the null space of the Levi form. For this, positivity of the Hessian of a defining function at a boundary point is needed only on the span of the null space of the Levi form and the complex normal. The situation was cleared up in [92]: the authors showed that the vector fields and the plurisubharmonic defining functions approaches can be reformulated naturally and then become equivalent.

Let Ω be a smooth bounded pseudoconvex domain. Say that Ω admits a family of essentially pluriharmonic defining functions if there exists a family $\{\rho_\varepsilon\}_{\varepsilon>0}$ of defining functions with gradients bounded and bounded away from zero on $b\Omega$ uniformly in ε, such that the complex Hessian of ρ_ε is $O(\varepsilon)$ on the span, over $\mathbb{C}$, of $\mathcal{N}_P$ and $L_n(P)$, for all $P \in b\Omega$. We emphasize that this notion is indeed a generalization of the notion of a pluri*sub*harmonic defining function (see [92]). We say that the form α (see above) is approximately exact on the null space of the Levi form if there exists a family $\{h_\varepsilon\}_{\varepsilon>0}$ of functions smooth in neighborhoods U_ε of the set K of boundary points of infinite D'Angelo type, bounded uniformly in ε, such that $dh|_{\mathcal{N}_P} = \alpha|_{\mathcal{N}_P} + O(\varepsilon)$ for all $P \in K$. A family of conjugate normals which are approximately holomorphic in weakly pseudoconvex directions is defined similarly; see [92], where Sucheston and the author established the following equivalence (compare also the remarks in section 5 of [91] concerning (iii)).

Theorem 3.4. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^n$. The following are equivalent:*

(i) *Ω admits a family of essentially pluriharmonic defining functions.*

(ii) *Ω admits a family of conjugate normals which are approximately holomorphic in weakly pseudoconvex directions.*

(iii) *Ω admits a family of vector fields as in* [13].

(iv) *The form α is approximately exact on the null space of the Levi form.*

The equivalence to condition (ii) is of interest because the existence of such a family leads, under favorable circumstances (K is uniformly H-convex), to the existence of transverse vector fields *holomorphic in a neighborhood of K*, and these lead to Stein neighborhood bases for $\overline{\Omega}$ and to Mergelyan type approximation ([44], [92], [45]).

Theorem 3.4 shows that the approaches to global regularity in the $\bar\partial$-Neumann problem through plurisubharmonic defining functions and through good vector fields

are really equivalent. Left unanswered was the question of how to unify this approach with that via compactness. This is achieved in the following theorem from [91].

Theorem 3.5. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^n$, ρ a defining function for Ω. Let $1 \leq q \leq n$. Assume that there is a constant C such that for all $\varepsilon > 0$ there exists a defining function ρ_ε for Ω and a constant C_ε with*

$$1/C < |\nabla \rho_\varepsilon| < C \quad \text{on } b\Omega, \tag{3.5}$$

and

$$\Big\| \sum_{|K|=q-1}{}' \Big(\sum_{j,k=1}^{n} \frac{\partial^2 \rho_\varepsilon}{\partial z_j \partial \overline{z}_k} \frac{\partial \rho}{\partial \overline{z}_j} \bar{u}_{kK} \Big) d\overline{z}_K \Big\|^2 \leq \varepsilon \big(\|\bar{\partial} u\|^2 + \|\bar{\partial}^* u\|^2 \big) + C_\varepsilon \|u\|_{-1}^2 \tag{3.6}$$

for all $u \in C^\infty_{(0,q)}(\overline{\Omega}) \cap \mathrm{Dom}(\bar{\partial}^*)$. Then

$$\|N_q u\|_s \leq C_s \|u\|_s, \tag{3.7}$$

for $s \geq 0$ and all $u \in W^s_{(0,q)}(\Omega)$.

Notice that the assumptions in Theorem 3.5 are for q-forms, q fixed. It is not hard to see that when they are satisfied at level q, then they are satisfied at level $q+1$ ([91], Lemma 2). It would be interesting to know whether global regularity similarly moves up to higher form levels (recall from Section 2 that subellipticity and compactness do).

The simplest situation occurs when there is one defining function, say ρ, that works for all ε. This covers the case when N_q is compact: the left hand side is in this case bounded by $\|u\|^2$ independently of ε, and compactness says precisely that $\|u\|^2$ can be bounded in the manner required by the right hand side of (3.6) (this is the right hand side of a compactness estimate).

When Ω admits a defining function ρ that is plurisubharmonic at the boundary, $\rho_\varepsilon = \rho$ for all ε also works. Assume $q = 1$ for the moment. Applying the Cauchy-Schwarz inequality to the left hand side of (3.6) at boundary points gives that this left hand side is dominated by $\sum(\partial^2 \rho / \partial z_j \partial \overline{z}_k) u_j \bar{u}_k$ plus a term of order ρ plus a compactly supported term. The latter two are benign for (3.6). Estimating the former in the way required in (3.6) can be done via subelliptic multiplier properties of the Levi matrix ([34], Lemma 4.1), or via the Kohn–Morrey formula (see [91] for details). When $q > 1$, one can reformulate (3.6) so that the left hand side of the inequality involves a pairing between q-forms ([91], Lemma 1), and the above argument works under the weaker assumption that the sum of any q eigenvalues of the Hessian of ρ is nonnegative. In view of the equivalence results in [10], this recovers, in the pseudoconvex case, a recent result of Herbig–McNeal ([56]), where the authors prove Sobolev estimates for the Bergman projection on j-forms, $q-1 \leq j \leq n$, under this weaker assumption.

More generally, the sufficient conditions for global regularity from Theorem 3.4 imply those in Theorem 3.5:

Proposition 3.6. *Let Ω be a smooth bounded pseudoconvex domain in $\mathbb{C}^n$. Assume that Ω satisfies one (hence all) of the equivalent conditions in Theorem* 3.4. *Then the assumptions in Theorem* 3.5 *are satisfied for $q = 1, 2, \ldots, n$.*

We indicate what is involved, details are in [91], Proposition 1. Assume (i) in Theorem 3.4. It suffices to consider the case $q = 1$. Fix ε. Then $L_{\rho_\varepsilon}(z)(w) = O(\varepsilon)|w|^2$ when (z, w) is in a neighborhood U_ε of the compact subset $\{(z, w) : w \in \mathcal{N}_z\}$ of the unit sphere bundle in $T^{1,0}(b\Omega)$. Here, L_g denotes the complex Hessian of a function g. There is a constant C_ε such that $|w|^2 \le C_\varepsilon L_{\rho_\varepsilon}(z)(w)$ when $(z, w) \notin U_\varepsilon$. This implies the estimate, when $z \in b\Omega$, $w \in T^{1,0}(b\Omega)$:

$$\Big| \sum_{j,k=1}^{n} \frac{\partial^2 \rho_\varepsilon}{\partial z_j \partial \bar z_k}(z) \frac{\partial \rho}{\partial \bar z_j}(z) \bar w_k \Big|^2 \le C\varepsilon |w|^2 + \widetilde{C_\varepsilon} \sum_{j,k=1}^{n} \frac{\partial^2 \rho_\varepsilon}{\partial z_j \partial \bar z_k}(z) w_j \bar w_k \tag{3.8}$$

(since both terms on the right are nonnegative). By continuity and homogeneity, (3.8) holds near (depending on ε) the boundary. To verify (3.6) for $u \in C^\infty_{(0,1)}(\overline{\Omega})$, it suffices to apply (3.8) to u pointwise, near the boundary (the normal component of u is zero only on the boundary, but it satisfies a subelliptic estimate, so is under control). In view of the discussion preceding the statement of Proposition 3.6, integration over Ω now gives (3.6).

It should not be surprising that condition (3.6) has a potential theoretic flavor: global regularity probably is not determined by geometric conditions alone (unlike the much stronger property of subellipticity). However, it is not hard to extract a geometric sufficient condition from (3.6), compare [91], section 2. What one arrives at is precisely condition (i) in Theorem 3.4. In other words, *the vector field approach constitutes* what might be called *the geometric content of Theorem* 3.5.

It is noteworthy that whether or not a family of defining functions satisfies (3.5) and (3.6) is determined entirely by the interplay of the gradients with the boundary. That is, if a family $\{\rho_\varepsilon\}_{\varepsilon>0}$ satisfies (3.5) and (3.6), and $\{\widetilde{\rho_\varepsilon}\}_{\varepsilon>0}$ is another family such that $\nabla(\widetilde{\rho_\varepsilon}) = \nabla(\rho_\varepsilon)$ for all ε and all $z \in b\Omega$, then $\{\widetilde{\rho_\varepsilon}\}_{\varepsilon>0}$ also satisfies (3.5) and (3.6) (possibly after rescaling). For details, see [91], Remark 2.

At the level of a priori estimates, a proof of Theorem 3.5 follows from a small modification of the ideas in [11], [13]. We briefly indicate what changes, keeping the general setup from [11]. This will show how (3.6) enters into the estimates. Set $X_\varepsilon = e^{-h_\varepsilon} \sum (\partial\rho/\partial\bar z_j)(\partial/\partial z_j)$, where h_ε is defined by $\rho_\varepsilon = e^{h_\varepsilon}\rho$, and ρ is a defining function with normalized gradient (all of this is near $b\Omega$, away from $b\Omega$, any smooth continuation will do). For the Bergman projection P, the key quantity to be estimated is

$$\|\varphi(X_\varepsilon - \overline{X}_\varepsilon)Pf\|^2 \lesssim \big(N_1 \bar\partial f, \varphi^2 (X_\varepsilon - \overline{X}_\varepsilon)[\bar\partial, X_\varepsilon - \overline{X}_\varepsilon]Pf\big) + \text{o.k.}, \tag{3.9}$$

where 'o.k.' stands for terms that are under control or can be absorbed, and φ is a smooth cutoff function supported near the boundary (see [11], p. 83–84). One needs to control the normal component of the commutator in (3.9). In contrast to [11], we

do not have a pointwise estimate on this normal component. But there is some slack built into the argument in [11], in that there the contribution from the commutator of $X_\varepsilon - \overline{X}_\varepsilon$ with each component of $\bar\partial$ is estimated separately. If one takes this into account, computing the commutator gives the main term (after integrating $X_\varepsilon - \overline{X}_\varepsilon$ back to the left)

$$\Big(\sum_{j,k=1}^{n} \frac{\partial^2 \rho_\varepsilon}{\partial z_j \partial \overline{z}_k} \big((X_\varepsilon - \overline{X}_\varepsilon) N_1 \bar\partial f \big)_j \frac{\partial \rho}{\partial z_k},\ X_\varepsilon P f \Big) \tag{3.10}$$

(as opposed to estimating $\sum_{k=1}^n \cdots$ for each j, $j = 1, \ldots, n$). The term in the left hand side of this inner product is now (the conjugate of) one to which (3.6) can be applied. (As usual, we let $X_\varepsilon - \overline{X}_\varepsilon$ act in special boundary charts so that it preserves $\mathrm{Dom}(\bar\partial^*)$.) Note that $X_\varepsilon = e^{h_\varepsilon} L_n$, and that e^{h_ε} is bounded independently of ε. Consequently (by (3.6)), the square of the L^2-norm of this term is dominated by

$$\begin{aligned} \varepsilon \big(\|\bar\partial (L_n - \overline{L_n}) N_1 \bar\partial f\|^2 + \|\bar\partial^* (L_n - \overline{L_n}) N_1 \bar\partial f\|^2 \big) + C_\varepsilon \|(L_n - \overline{L_n}) N_1 \bar\partial f\|_{-1}^2 \\ \lesssim \varepsilon \big(\|N_1 \bar\partial f\|_1^2 + \|\bar\partial^* N_1 \bar\partial f\|_1^2 \big) + C_\varepsilon \|f\|^2. \end{aligned} \tag{3.11}$$

From here on, the argument proceeds as in [11]; in particular, in the setup of the downward induction on q there, $N_1 \bar\partial$ is 'as good as' P. Absorbing terms, one arrives at the required a priori estimate (compare p. 84–85 in [11]).

The situation changes rather markedly with regard to genuine estimates. In [11], the authors simply observe that the estimates can be carried out uniformly on suitable approximating strictly pseudoconvex subdomains, by using the same family of vector fields. By contrast, the assumptions in Theorem 3.5 do not seem strong enough to be inherited by these approximating subdomains. Therefore, one has to employ some other regularization procedure, such as elliptic regularization. This makes the argument considerably more involved, and the author derives in [91] certain needed new estimates for the regularized operators. This is also in contrast to [23], where the results of [11] are proved working directly with the $\bar\partial$-Neumann operator and using elliptic regularization. There too it is the strength of the pointwise estimates on the size of the normal component of the commutators that makes elliptic regularization routine (once the derivation of the a priori estimates is in place).

Note that to get estimates at a fixed Sobolev level k, it suffices to have (3.6) in Theorem 3.5 for some $\varepsilon = \varepsilon(k)$. Kohn ([68]) has proved estimates where the level in the Sobolev scale up to which estimates hold is tied to the Diederich–Fornæss exponent ([39]) of the domain. The discussion above of Proposition 3.6, combined with [89], where the plurisubharmonicity of $-\log(-\rho)$ is exploited, suggests that it should be possible to obtain results of this type by the methods in [91].

Remark 3.7. Consider the operator A_ρ from $\mathrm{Dom}(\bar\partial) \cap \mathrm{Dom}(\bar\partial^*)$, provided with the graph norm, to $L^2(\Omega)$, given by

$$A_\rho(u) = \sum_{j,k=1}^{n} \frac{\partial^2 \rho}{\partial z_j \partial \bar z_k} \frac{\partial \rho}{\partial \bar z_j} \bar u_k, \quad u \in \mathrm{Dom}(\bar\partial) \cap \mathrm{Dom}(\bar\partial^*). \tag{3.12}$$

Then (3.6) holds with $\rho_\varepsilon = \rho$ for all ε precisely when A_ρ is compact (see e.g. [22], Lemma 1, [75], Lemma 2.1). The form of A_ρ suggests that one study sesquilinear forms that produce compact operators via (3.12). It is possible that there is a theory of 'compactness multipliers'. We mention that compactness of A_ρ for a suitable defining function ρ is considerably weaker than compactness of N_1. It holds on all convex domains (since they admit a plurisubharmonic defining function), yet N_1 is compact (if and) only if the boundary of the domain contains no analytic disc ([46]).

References

[1] Andreotti, A., and Vesentini, E., Carleman estimates for the Laplace-Beltrami equations on complex manifolds. *Inst. Hautes Études Sci. Publ. Math.* **25** (1965), 81–130.

[2] Barrett, David E., Irregularity of the Bergman projection on a smooth bounded domain in $\mathbb{C}^2$. *Ann. of Math.* (2) **119** (1984), 431–436.

[3] —, Behavior of the Bergman projection on the Diederich-Fornæss worm. *Acta Math.* **168** (1992), 1–10.

[4] Barrett, D. E., and Fornæss, J. E., On the smoothness of Levi-foliations. *Publ. Mat.* **32**, Nr.2 (1988), 171–177.

[5] Bedford, Eric, and Fornæss, John Erik, Domains with pseudoconvex neighborhood systems. *Invent. Math.* **47** (1978), 1–27.

[6] Bell, S., Mapping problems in complex analysis and the $\bar\partial$-problem. *Bull. Amer. Math. Soc.* (N.S.) **22** (2) (1990), 233–259.

[7] Bell, S., and Catlin, D., Boundary regularity of proper holomorphic mappings. *Duke Math. J.* **49** (1982), 385–396.

[8] Berndtsson, B., $\bar\partial$ and Schrödinger operators. *Math. Z.* **221** (1996), 401–413.

[9] —, Weighted estimates for the $\bar\partial$-equation. In *Complex Analysis and Geometry* (ed. by Jeffery McNeal), Ohio State Univ. Math. Res. Inst. Publ. 9, Walter de Gruyter, Berlin 2001, 141–160.

[10] Boas, Harold P., and Straube, Emil J., Equivalence of regularity for the Bergman projection and the $\bar\partial$-Neumann operator. *Manuscripta Math.* **67** (1990), 25–33.

[11] —, Sobolev estimates for the $\bar\partial$-Neumann operator on domains in $\mathbb{C}^n$ admitting a defining function that is plurisubharmonic on the boundary. *Math. Z.* **206** (1991), 81–88.

[12] —, The Bergman projection on Hartogs domains in $\mathbb{C}^2$. *Trans. Amer. Math. Soc.* **331** (1992), 529–540.

[13] —, De Rham cohomology of manifolds containing the points of infinite type, and Sobolev estimates for the $\bar\partial$-Neumann problem. *J. Geom. Anal.* **3** (3) (1993), 225–235.

[14] —, Global regularity of the $\bar\partial$-Neumann problem: a survey of the L^2-Sobolev theory. In *Several Complex Variables* (ed. by M. Schneider and Y.-T. Siu), Cambridge University Press, Cambridge 1999, 79–111.

[15] Candel, Alberto, and Conlon, Lawrence, *Foliations I*. Grad. Stud. Math. 23, Amer. Math. Soc., Providence, RI, 2000.

[16] Catlin, David W., Boundary behavior of holomorphic functions on weakly pseudoconvex domains. Dissertation, Princeton University, 1978.

[17] —, Necessary conditions for subellipticity and hypoellipticity for the $\bar{\partial}$-Neumann problem on pseudoconvex domains. In *Recent Developments in Several Complex Variables* (ed. by John E. Fornæss), Ann. of Math. Stud. 100, Princeton University Press, Princeton, N.J., 1981, 93–100.

[18] —, Necessary conditions for subellipticity of the $\bar{\partial}$-Neumann problem. *Ann. of Math.* (2) **117** (1983), 147–171.

[19] —, Boundary invariants of pseudoconvex domains. *Ann. of Math.* (2) **120** (1984), 529–586.

[20] —, Global regularity of the $\bar{\partial}$-Neumann problem. In *Complex Analysis of Several Variables* (ed. by Y.-T. Siu), Proc. Sympos. Pure Math. 41, Amer. Math. Soc., Providence, RI, 1984, 39–49.

[21] —, Subelliptic estimates for the $\bar{\partial}$-Neumann problem. *Ann. of Math.* (2), **126** (1987), 131–191.

[22] Catlin, David W., and D'Angelo, John P., Positivity conditions for bihomogeneous polynomials. *Math. Res. Lett.* **4** (4) (1997), 555–567.

[23] Chen, So-Chin, and Shaw, Mei-Chi, *Partial Differential Equations in Several Complex Variables*. AMS/IP Stud. Adv. Math. 19, Amer. Math. Soc./International Press, Providence, RI/Boston, MA, 2001.

[24] Christ, Michael, Global C^∞ irregularity of the $\bar{\partial}$-Neumann problem for worm domains. *J. Amer. Math. Soc.* **9** (4) (1996), 1171–1185.

[25] —, Singularity and regularity – local and global. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 627–636.

[26] —, Remarks on global irregularity in the $\bar{\partial}$-Neumann problem. In *Several Complex Variables* (ed. by M. Schneider and Y.-T. Siu), Cambridge University Press, Cambridge 1999, 161–198.

[27] Christ, Michael, and Fu, Siqi, Compactness of the $\bar{\partial}$-Neumann problem, magnetic Schrödinger operators, and the Aharonov-Bohm effect. *Adv. Math.* **197** (1) (2005), 1–40.

[28] D'Angelo, John P., Finite type conditions for real hypersurfaces. *J. Diff. Geometry* **14** (1979), 59–66.

[29] —, Subelliptic estimates and failure of semicontinuity of orders of contact. *Duke Math. J.* **47** (1980), 955–957.

[30] —, Real hypersurfaces, orders of contact, and applications. *Ann. of Math.* (2) **115** (1982), 615–637.

[31] —, Iterated commutators and derivatives of the Levi form. In *Complex Analysis* (ed. by Steven G. Krantz), Lecture Notes in Math. 1268, Springer-Verlag, Berlin 1987, 103–110.

[32] —, *Several Complex Variables and the Geometry of Real Hypersurfaces*. Stud. Adv. Math., CRC Press, Boca Raton, FL, 1993.

[33] —, A gentle introduction to points of finite type on real hypersurfaces. In *Explorations in Complex and Riemannian Geometry* (ed. by John Bland et al.), Contemp. Math. 332, Amer. Math. Soc., Providence, RI, 2003, 19–36.

[34] D'Angelo, John P., and Kohn, Joseph J., Subelliptic estimates and finite type. In *Several Complex Variables* (ed. by M. Schneider and Y.-T. Siu), Cambridge University Press, Cambridge 1999, 199–232.

[35] Demailly, Jean-Pierre, Multiplier ideal sheaves and analytic methods in algebraic geometry. In *School on Vanishing Theorems and Effective Results in Algebraic Geometry* (Trieste, 2000), ICTP Lect. Notes 6, Abdus Salam Int. Cent. Theoret. Phys., Trieste 2001, 1–148.

[36] —, L^2 Hodge theory and vanishing theorems. In *Introduction to Hodge Theory*, by José Bertin, Jean-Pierre Demailly, Luc Illusie, and Chris Peters, SMF/AMS Texts Monogr. 8, Amer. Math. Soc./ Soc. Math. France, Providence, RI/Paris 2002, 1–98.

[37] Derridj, M., Regularité pour $\overline{\partial}$ dans quelques domaines faiblement pseudoconvexes. *J. Differential Geom.* **13** (1978), 559–576.

[38] Diederich, K., and Fornæss, J. E., Pseudoconvex domains: an example with nontrivial Nebenhülle. *Math. Ann.* **255** (1977), 275–292.

[39] —, Pseudoconvex domains: bounded strictly plurisubharmonic exhaustion functions. *Invent. Math.* **39** (1977), 129–141.

[40] —, Pseudoconvex domains with real-analytic boundary. *Ann. of Math.* (2) **107** (2) (1978), 371–384

[41] —, Boundary regularity of proper holomorphic mappings. *Invent. Math.* **67** (1982), 363–384.

[42] Diederich, Klas, and Pflug, Peter, Necessary conditions for hypoellipticity of the $\overline{\partial}$-problem. In *Recent Developments in Several Complex Variables* (ed. by John E. Fornæss), Ann. of Math. Stud. 100, Princeton University Press, Princeton, N.J., 1981, 151–154.

[43] Folland, G. B., and Kohn, J. J., *The Neumann Problem for the Cauchy-Riemann Complex*. Ann. of Math. Stud. 75, Princeton University Press, Princeton, N.J., 1972.

[44] Fornæss, John Erik, and Nagel, Alexander, The Mergelyan property for weakly pseudoconvex domains. *Manuscripta Math.* **22** (1977), 199–208.

[45] Forstnerič, Franc, and Laurent-Thiébaut, Christine, Stein compacts in Levi-flat hypersurfaces. *Trans. Amer. Math. Soc.*, to appear.

[46] Fu, Siqi, and Straube, Emil J., Compactness of the $\overline{\partial}$-Neumann problem on convex domains. *J. Funct. Anal.* **159** (1998), 629–641.

[47] —, Compactness in the $\overline{\partial}$-Neumann problem. In *Complex Analysis and Geometry* (ed. by J. McNeal), Ohio State Univ. Math. Res. Inst. Publ. 9, Walter de Gruyter, Berlin 2001, 141–160.

[48] —, Semi-classical analysis of Schrödinger operators and compactness in the $\overline{\partial}$-Neumann problem. *J. Math. Anal. Appl.* **271** (2002), 267–282; Correction *ibid* **280** (2003), 195–196.

[49] Fuglede, B., The Dirichlet Laplacian on finely open sets. *Potential Anal.* **10** (1999), 91–101.

[50] Harrington, Phillip S., A quantitative analysis of Oka's lemma. Preprint.

[51] Hefer, Torsten, and Lieb, Ingo, On the compactness of the $\overline{\partial}$-Neumann operator. *Ann. Fac. Sci. Toulouse Math.* (6) **9**, Nr.3 (2000), 415–432.

[52] Helffer, B., *Semi-Classical Analysis for the Schrödinger Operator and Applications*. Lecture Notes in Math. 1336, Springer-Verlag, Berlin 1988.

[53] Helms, L. L., *Introduction to Potential Theory*. Pure Appl. Math. 22, Wiley-Interscience, New York, London, Sydney 1969.

[54] Henkin, G. M., and Iordan, A., Compactness of the Neumann operator for hyperconvex domains with non-smooth B-regular boundary. *Math. Ann.* **307** (1997), 151–168.

[55] Herbig, Anne-Katrin, A sufficient condition for subellipticity of the $\bar\partial$-Neumann operator. Preprint.

[56] Herbig, Anne-Katrin, and McNeal, Jeffery D., Regularity of the Bergman projection on forms and plurisubharmonicity conditions. *Math. Ann.*, to appear.

[57] Hodge, W. V. D., *The Theory and Applications of Harmonic Integrals*. Cambridge University Press, Cambridge 1941.

[58] Hörmander, L., L^2 estimates and existence theorems for the $\bar\partial$ operator. *Acta Math.* **113** (1965), 89–152.

[59] —, *An Introduction to Complex Analysis in Several Variables*. Third ed., North-Holland, Amsterdam 1990.

[60] —, A history of existence theorems for the Cauchy-Riemann complex in L^2 spaces. *J. Geom. Anal.* **13**, (2) (2003), 329–357.

[61] Kato, T., Schrödinger operators with singular potentials. *Israel J. Math.* **13** (1972), 135–148.

[62] Kim, Mijoung, The $\bar\partial$-Neumann operator and the Kobayashi metric. *Illinois J. Math.* **48** (2) (2004), 635–643.

[63] Kiselman, C. O., A study of the Bergman projection in certain Hartogs domains. In *Several Complex Variables and Complex Geometry* (ed. by E. Bedford et al.), Proc Sympos. Pure Math. 52, Amer. Math. Soc., Providence, RI, 1991, 219–231.

[64] Kohn, J. J., Solution of the $\bar\partial$-Neumann problem on strongly pseudo-convex manifolds. *Proc. Nat. Acad. Sci. USA* **47** (1961), 1198–1202.

[65] —, Harmonic integrals on strongly pseudoconvex manifolds, I. *Ann. of Math.* (2) **78** (1963), 112–148.

[66] —, Harmonic integrals on strongly pseudoconvex manifolds, II. *Ann. of Math.* (2) **79** (1964), 450–472.

[67] —, Subellipticity of the $\bar\partial$-Neumann problem on pseudoconvex domains: sufficient conditions. *Acta Math.* **142** (1979), 79–122.

[68] —, Quantitative estimates for global regularity. In *Analysis and Geometry in Several Complex Variables* (ed. by G. Komatsu and M. Kuranishi), Trends Math. Birkhäuser, Boston, MA, 1999, 97–128.

[69] —, Contribution in Donald C. Spencer (1912–2001), *Notices Amer. Math. Soc.* **51**, (1) (2004), 17–29.

[70] —, Ideals of multipliers. In *Complex Analysis in Several Variables—Memorial Conference of Kiyoshi Oka's Centennial Birthday*, Adv. Stud. Pure Math. 42, Math. Soc. Japan, Tokyo 2004, 147–157.

[71] Kohn, J. J., and Nirenberg, L., Non-coercive boundary value problems. *Comm. Pure Appl. Math.* **18** (1965), 443–492.

[72] Krantz, Steven G., *Partial Differential Equations and Complex Analysis*. Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[73] Lieb, Ingo, and Michel, Joachim, *The Cauchy-Riemann Complex. Integral Formulae and Neumann Problem*. Aspects Math. E34, Vieweg, Wiesbaden 2002.

[74] Matheos, Peter, A Hartogs domain with no analytic discs in the boundary for which the $\bar\partial$-Neumann problem is not compact. Dissertation, University of California, Los Angeles, 1997.

[75] McNeal, Jeffery D., A sufficient condition for compactness of the $\bar\partial$-Neumann operator. *J. Funct. Anal.* **195** (2002), Nr. 1, 190–205.

[76] —, Private communication

[77] —, $\mathcal{L}^2$ estimates on twisted Cauchy-Riemann complexes, Cont. Math., sesquicentennial volume for Washington University, 2005

[78] Morrey, C. B., Jr., The analytic embedding of abstract real-analytic manifolds. *Ann. of Math.* (2) **68** (1958), 159–201.

[79] —, The $\bar\partial$-Neumann problem on strongly pseudoconvex manifolds. In *Differential Analysis*, Tata Institute of Fundamental Research Studies in Mathematics 2, Oxford University Press, London 1964, 81–133.

[80] Munasinghe, Samangi, Dissertation, Texas A&M University, in preparation

[81] Ohsawa, Takeo, *Analysis of Several Complex Variables*. Transl. Math. Monogr. 211, Amer. Math. Soc., Providence, RI, 2002.

[82] Şahutoğlu, Sönmez, and Straube, Emil J., Analytic discs, plurisubharmonic hulls, and non-compactness of the $\bar\partial$-Neumann operator. *Math. Ann.* **334** (2006), 809–820.

[83] Shaw, Mei-Chi, Boundary value problems on Lipschitz domains in $\mathbb{R}^n$ or $\mathbb{C}^n$. In *Geometric Analysis of Partial Differential Equations and Several Complex Variables*, Contemp. Math. 368, Amer. Math. Soc., Providence, RI, 2005, 375–404.

[84] Sibony, Nessim, Une classe de domaines pseudoconvexes. *Duke Math. J.* **55** , Nr. 2 (1987), 299–319.

[85] Simon, B., Universal diamagnetism of spinless boson systems. *Phys. Rev. Lett.* **36** (1976), 804–806.

[86] Siu, Yum-Tong, Some recent transcendental techniques in algebraic and complex geometry. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. I, Higher Ed. Press, Beijing 2002, 439–448.

[87] —, Multiplier ideal sheaves in complex and algebraic geometry. *Sci. China Ser. A* **48** (2005), 1–31.

[88] Straube, Emil J., Plurisubharmonic functions and subellipticity of the $\bar\partial$-Neumann problem on non-smooth domains. *Math. Res. Lett.* **4** (1997), 459–467.

[89] —, Good Stein neighborhood bases and regularity of the $\bar\partial$-Neumann problem. *Illinois J. Math.* **45** (2001), 865–871.

[90] —, Geometric conditions which imply compactness of the $\bar\partial$-Neumann operator. *Ann. Inst. Fourier Grenoble* **54** (3) (2004), 699–710.

[91] —, A sufficient condition for global regularity of the $\bar\partial$-Neumann operator. Preprint.

[92] Straube, Emil J., and Sucheston, Marcel K., Plurisubharmonic defining functions, good vector fields, and exactness of a certain one form. *Monatsh. Math.* **136** (2002), 249–258.

[93] —, Levi foliations in pseudoconvex boundaries and vector fields that commute approximately with $\bar\partial$. *Trans. Amer. Math. Soc.* **355** (1) (2003), 143–154.

[94] Sullivan, Dennis, Cycles for the dynamical study of foliated manifolds and complex manifolds. *Invent. Math.* **36** (1976), 225–255.

[95] Takegoshi, Kensho, A new method to introduce a priori estimates for the $\bar{\partial}$-Neumann problem. In *Complex Analysis* (ed. by K. Diederich), Aspects Math. E17, Vieweg, Wiesbaden 1991, 310–314.

[96] Tondeur, Philippe, *Geometry of Foliations*. Monogr. Math. 90, Birkhäuser, Basel 1997.

Department of Mathematics, Texas A&M University, College Station, TX 77843, U.S.A.
E-mail: straube@math.tamu.edu

Greedy approximations with regard to bases

Vladimir N. Temlyakov

Abstract. This paper is a survey of recent results on greedy approximations with regard to bases. The theory of greedy approximations is a part of nonlinear approximations. The standard problem in this regard is the problem of m-term approximation where one fixes a basis and seeks to approximate a target function by a linear combination of m terms of the basis. When the basis is a wavelet basis or a basis of other waveforms, then this type of approximation is the starting point for compression algorithms. We are interested in the quantitative aspects of this type of approximation. Introducing the concept of best m-term approximation we obtain a lower bound for the accuracy of any method providing m-term approximation. It is known that a problem of simultaneous optimization over many parameters (like in best m-term approximation) is a very difficult problem. We would like to have an algorithm for constructing m-term approximants that adds at each step only one new element from the basis and keeps elements of the basis obtained at the previous steps. The primary object of our discussion is the Thresholding Greedy Algorithm (TGA) with regard to a given basis. The TGA, applied to a function f, picks at the mth step an element with the mth biggest coefficient (in absolute value) of the expansion of f in the series with respect to the basis. We show that this algorithm is very good for a wavelet basis and is not that good for the trigonometric system. We discuss in detail the behavior of the TGA with regard to the trigonometric system. We also discuss one example of an algorithm from a family of very general greedy algorithms that works in the case of a redundant system instead of a basis. It turns out that this general greedy algorithm is very good for the trigonometric system.

Mathematics Subject Classification (2000). Primary 41A25; Secondary 41A46.

Keywords. Nonlinear approximation, greedy algorithm, convergence, best m-term approximation, greedy basis.

1. Introduction. Historical remarks

In order to give the reader some ideas for comparing the quality of approximation methods we now discuss some classical results in approximation of periodic functions. In this section we briefly discuss various classical approaches, created in linear approximation, for the estimation of the quality of a method of approximation. We will use and refine these approaches in nonlinear approximation. We confine our discussion to the case of approximation of periodic functions of a single variable. The two main parameters of a method of approximation are accuracy and complexity. These concepts may be treated in various ways depending on the particular problems involved. Here we will start from the classical idea about approximation of functions by polynomials. After Fourier's article (1807) the representation of a 2π-periodic

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

function by its Fourier series became natural. In other words, the function $f(x)$ is approximately represented by a partial sum $S_n(f,x)$ of its Fourier series.

We will be interested in the approximation of a function f by a polynomial $S_n(f)$ in some L_p-norm, $1 \le p \le \infty$. In the case $p = \infty$ we will assume that we deal with the uniform norm. As accuracy of the method of approximating a periodic function by its Fourier partial sum we will consider the quantity $\|f - S(f)\|_p$. The complexity of this method of approximation contains the two following characteristics. The order of the trigonometric polynomial $S_n(f)$ is the quantitative characteristic. The following observation gives us the qualitative characteristic. The coefficients of this polynomial are found by the Fourier formulas which means that the operator S_n is the orthogonal projection onto the subspace of trigonometric polynomials of order n.

In 1854 Chebyshev suggested to represent a continuous function f by its polynomial of best approximation, namely by the polynomial $t_n(f)$ such that

$$\|f - t_n(f)\|_\infty = E_n(f)_\infty := \inf_{\alpha_k,\beta_k} \left\| f(x) - \sum_{k=0}^{n} (\alpha_k \cos kx + \beta_k \sin kx) \right\|_\infty .$$

He proved the existence and uniqueness of such a polynomial. We will consider this method of approximation not only in the uniform norm, but in all L_p-norms, $1 \le p < \infty$. The accuracy of the Chebyshev method can be easily compared with the accuracy of the Fourier method:

$$E_n(f)_p \le \|f - S_n(f)\|_p.$$

The quantitative characteristics of complexity coincide for the two methods but the qualitative characteristics are different (for example, it is not difficult to understand that for $p = \infty$ the mapping $f \to t_n(f)$ is not a linear operator). The du Bois-Reymond example (1873) of a continuous function f such that $\|f - S_n(f)\|_\infty \to \infty$ when $n \to \infty$, and the Weierstrass theorem which says that for each continuous function f we have $E_n(f)_\infty \to 0$ as $n \to \infty$, showed the advantage of the Chebyshev method in comparison with the Fourier method from the point of view of accuracy. It is known that for each $f \in L_2(\mathbb{T})$ the approximation with the error $E_n(f)_2$ can be realized by the operator S_n of orthogonal projection onto the space of trigonometric polynomials of order n. The performance of the operator S_n was studied thoroughly in all L_p spaces, $1 \le p \le \infty$. It was proved that S_n provides almost optimal or close to optimal approximation for each $f \in L_p(\mathbb{T})$:

$$\|f - S_n(f)\|_p \le C(p) E_n(f)_p, \qquad 1 < p < \infty,$$
$$\|f - S_n(f)\|_p \le C \ln(n+2) E_n(f)_p, \quad p = 1, \infty.$$

The desire to construct methods of approximation which have the advantages of the Fourier and Chebyshev methods led to the study of various methods of summation of the Fourier series. The most important among them from the point of view of approximation are the de la Vallée Poussin, Fejér and Jackson methods which were

constructed early in the 20th century. All these methods are linear. For example, the de la Vallée Poussin method is the method of approximation of a function f by the polynomial

$$V_n(f) = \frac{1}{n} \sum_{l=n}^{2n-1} S_l(f)$$

of order $2n - 1$.

From the point of view of accuracy this method is close to the Chebyshev method; de la Vallée Poussin proved that

$$\|f - V_n(f)\|_p \le 4E_n(f)_p, \quad 1 \le p \le \infty.$$

From the point of view of complexity it is close to the Fourier method, and the property of linearity essentially distinguishes it from the Chebyshev method.

We see that common to all these methods is the approximation by means of trigonometric polynomials; however, the ways of constructing these polynomials differ: orthogonal projections on the subspace of trigonometric polynomials of fixed order, the operator of best approximation, and linear operators.

In 1936 Kolmogorov introduced the concept of width $d_n(F, X)$ of a class F in a Banach space X:

$$d_n(F, X) = \inf_{\{\phi_j\}_{j=1}^n} \sup_{f \in F} \inf_{\{c_j\}_{j=1}^n} \left\| f - \sum_{j=1}^n c_j \phi_j \right\|_X.$$

This concept is designed to find for a fixed n and for a class F a subspace of dimension n, optimal with respect to the construction of an approximating element as the element of best approximation. In other words, the Kolmogorov width gives the lower bound for accuracy of Chebyshev's methods, having the same quantitative characteristic of complexity (the dimension of the approximating subspace). In analogy to the concept of the Kolmogorov width, that is, to the problem concerning the best Chebyshev method, the problems concerning the best linear method and the best Fourier method were considered. Tikhomirov ([51]) introduced the concept of linear width

$$\lambda_n(F, X) = \inf_{A:\operatorname{rank} A \le n} \sup_{f \in F} \|f - Af\|_X.$$

The concept of orthowidth (Fourier width) was introduced in [38]:

$$\varphi_n(F, X) := d_n^{\perp}(F, X) := \inf_{\substack{\text{orthonormal} \\ \text{system } \{u_i\}_{i=1}^n}} \sup_{f \in F} \left\| f - \sum_{i=1}^n \langle f, u_i \rangle u_i \right\|_X.$$

We discuss these widths in more detail later in this section. We present here some well-known results for the Sobolev classes

$$W_q^r := \{f : f^{(r-1)} \text{ is absolutely continuous, } \|f^{(r)}\|_q \le 1\}.$$

The first result about widths, namely Kolmogorov's result (1936)

$$d_{2n+1}(W_2^r, L_2) = (n+1)^{-r},$$

showed that the best subspace of dimension $2n+1$ for approximation of classes of periodic functions is the subspace of trigonometric polynomials of order n. This result confirmed that the approximation of functions in the class W_2^r by trigonometric polynomials is natural. Further estimates of the widths $d_{2n+1}(W_q^r, L_p)$, $1 \le q$, $p \le \infty$, some of which are discussed here, showed that for some values of the parameters q, p the subspace of trigonometric polynomials of order n is optimal (in the sense of order) but for other values of q, p this subspace is not optimal.

The Ismagilov estimate [20] for the quantity $d_n(W_1^r, L_\infty)$ gave the first example where the subspace of trigonometric polynomials of order n is not optimal. This phenomenon was thoroughly studied by Kashin [22]. We remark that from the point of view of orthowidth the Fourier operator S_n is optimal (in the sense of order of approximation in the L_p-norm) for all Sobolev classes W_q^r with $1 \le q$, $p \le \infty$ with the exception of the two cases $q = p = 1$ and $q = p = \infty$.

All the above defined widths have as a starting point a function class F. Thus in this setting we choose a priori a function class F and look for optimal subspaces for approximation of a given class. The following results are well known [39]. We present these results for r a positive integer. In the case $q = 1$, $p = \infty$ we assume $r > 1$. For a number a we denote $(a)_+ := \max(a, 0)$.

A. In the case $1 \le p \le q \le \infty$ or $1 \le q \le p \le 2$ one has

$$\varphi_n(W_q^r, L_p) \asymp \lambda_n(W_q^r, L_p) \asymp d_n(W_q^r, L_p) \asymp n^{-r+(1/q-1/p)_+}. \tag{1.1}$$

B. In the case $1 \le q < p \le \infty$, $p > 2$, one has

$$d_n(W_q^r, L_p) \asymp n^{-r+(1/q-1/2)_+},$$
$$\lambda_n(W_q^r, L_p) \asymp n^{-r+\max(1/q-1/2,1/2-1/p)},$$
$$\varphi_n(W_q^r, L_p) \asymp n^{-r+1/q-1/p}.$$

In case A the classical trigonometric system provides the optimal orders for all widths, except for φ_n for $q = p = 1, \infty$. Let us discuss the more interesting case B for the particular choice $q = 2$ and $p = \infty$. We have

$$d_n(W_2^r, L_\infty) \asymp n^{-r}, \tag{1.2}$$
$$\lambda_n(W_2^r, L_\infty) \asymp \varphi_n(W_2^r, L_\infty) \asymp n^{-r+1/2}. \tag{1.3}$$

These relations show that if we drop the linearity requirement for the approximation method we gain in accuracy a factor $n^{-1/2}$. However, there is a big difficulty in realization of the estimate (1.2). We know by Kashin's result that there exists a subspace realizing (1.2) but we do not know a way to construct it. Thus it is only an existence theorem for now.

Let us discuss one more special case: $q = 1$ and $p = \infty$. In this case we have

$$d_n(W_1^r, L_\infty) \asymp \lambda_n(W_1^r, L_\infty) \asymp n^{-r+1/2} \tag{1.4}$$

and

$$\varphi_n(W_1^r, L_\infty) \asymp n^{-r+1}. \tag{1.5}$$

Therefore, by (1.4) the best possible approximation (in the sense of order) can be realized by a linear method, say, A_n. However, by (1.5) this linear method A_n is certainly not an orthogonal projector. Moreover, by [39] it cannot satisfy even the following much weaker restriction $\|A_n(e^{ikx})\|_2 \le C, k \in \mathbb{Z}$. This means that the optimal linear operator A_n is unstable. A small change in some of the Fourier coefficients of f may result in a big change of $\|A_n(f)\|_2$.

Let us make some conclusions now. In linear approximation of W_q^r in L_p the bottom line is given by $\varphi_n(W_q^r, L_p)$ where the approximation method is the simplest, namely orthogonal projection. Partial sums with regard to classical systems provide an optimal error of approximation for this width. The trigonometric system works for all $1 \le q, p \le \infty$ except for $(q, p) = (1, 1), (\infty, \infty)$. The wavelet systems (see [1]) work for all $1 \le q, p \le \infty$. In the example of the pair (W_1^r, L_∞) we have seen that we need to sacrifice important and convenient properties of the approximating operator in order to achieve better accuracy. In the example of (W_2^r, L_∞) we have seen that we need to pay even a bigger price for better accuracy in a form of proving only an existence theorem instead of providing a constructive method of approximation.

Our main interest in this paper is nonlinear approximation. We begin our discussion with the trigonometric system. Let $\mathcal{T}$ be the complex trigonometric system $\{e^{ikx}\}_{k\in\mathbb{Z}}$. Denote for $f \in L_p(\mathbb{T})$

$$\sigma_m(f, \mathcal{T})_p := \inf_{c_1,\dots,c_m;\phi_1,\dots,\phi_m\in\mathcal{T}} \left\| f - \sum_{j=1}^m c_j\phi_j \right\|_p$$

the best m-term trigonometric approximation of f in the L_p-norm. It is clear that one can get an upper estimate for $\sigma_{2m+1}(f, \mathcal{T})_p$ by approximating f by trigonometric polynomials of order m. Denote $\mathcal{T}(m)$ the subspace of trigonometric polynomials of order m and define

$$E_m(f, \mathcal{T})_p := \inf_{t\in\mathcal{T}(m)} \|f - t\|_p.$$

The first result that indicated an advantage of m-term approximation over approximation by trigonometric polynomials of order m is due to Ismagilov [20]:

$$\sigma_m(|\sin x|, \mathcal{T})_\infty \le C_\epsilon m^{-6/5+\epsilon}, \quad \text{for any } \epsilon > 0. \tag{1.6}$$

Let us compare it with the well-known result due to de la Vallée Poussin and Bernstein:

$$E_m(|\sin x|, \mathcal{T})_\infty \asymp m^{-1}. \tag{1.7}$$

Maiorov [35] improved the estimate (1.6):

$$\sigma_m(|\sin x|, \mathcal{T})_\infty \asymp m^{-3/2}. \tag{1.8}$$

In [11] we proved the following rate of best m-term approximation of the Sobolev classes W_q^r in L_p, $1 \le q, p \le \infty$:

$$\sigma_m(W_q^r, \mathcal{T})_p := \sup_{f \in W_q^r} \sigma_m(f, \mathcal{T})_p \asymp m^{-r+(1/q-\max(1/p,1/2))_+}. \tag{1.9}$$

Comparing (1.9) with the above bounds for the Kolmogorov width we conclude that

$$\sigma_m(W_q^r, \mathcal{T})_p \asymp d_m(W_q^r, L_p).$$

In particular, this means, that in the case (W_2^r,L_∞) the nonlinear m-term approximations provide much better accuracy than the trigonometric polynomials of order m. The best m-term approximations $\sigma_m(f, \mathcal{T})_p$ may be considered as a nonlinear analogue (counterpart) of the best approximations $E_m(f, \mathcal{T})_p$. The main goal of this paper is to discuss a nonlinear analogue (counterpart) of the operator $S_n(f)$. We consider the greedy approximant to be a nonlinear analogue of the partial sum. In Sections 2 and 3 we discuss the general theory of greedy approximation with regard to bases. Our primary object of discussion is the Thresholding Greedy Algorithm (TGA). We return to a discussion of nonlinear approximations with regard to the trigonometric system in Sections 4 and 5. In Section 6 we deviate from the main stream of the paper of studying the TGA and discuss one example of a family of greedy algorithms that works in a very general situation. The most important feature of this algorithm is that it provides m-term approximation with regard to a very general system that may be redundant (overcomplete). It turns out (this will be seen from the discussion in Section 6) that this general approximation method is very good for the trigonometric system.

2. Greedy algorithms with regard to bases

Let X be a Banach space with a given basis $\Psi = \{\psi_k\}_{k=1}^\infty$. We assume that $\|\psi_k\| \ge C > 0$, $k = 1, 2, \dots$, and consider the following theoretical greedy algorithm. For a given element $f \in X$ we consider the expansion

$$f = \sum_{k=1}^\infty c_k(f, \Psi)\psi_k. \tag{2.1}$$

For an element $f \in X$ we call a permutation ρ, $\rho(j) = k_j$, $j = 1, 2, \dots$, of the positive integers decreasing and write $\rho \in D(f)$ if

$$|c_{k_1}(f, \Psi)| \ge |c_{k_2}(f, \Psi)| \ge \cdots. \tag{2.2}$$

In the case of strict inequalities here $D(f)$ consists of only one permutation. We define the m-th greedy approximant of f with regard to the basis Ψ corresponding to a permutation $\rho \in D(f)$ by the formula

$$G_m(f) := G_m(f, \Psi) := G_m(f, \Psi, \rho) := \sum_{j=1}^{m} c_{k_j}(f, \Psi)\psi_{k_j}.$$

We note that there is another natural greedy type algorithm based on ordering $\|c_k(f, \Psi)\psi_k\|$ instead of ordering absolute values of coefficients. In this case we do not need the restriction $\|\psi_k\| \geq C > 0$, $k = 1, 2, \ldots$. Denote by $\Lambda_m(f)$ a set of indices such that

$$\min_{k \in \Lambda_m(f)} \|c_k(f, \Psi)\psi_k\| \geq \max_{k \notin \Lambda_m(f)} \|c_k(f, \Psi)\psi_k\|.$$

We define $G_m^X(f, \Psi)$ by the formula

$$G_m^X(f, \Psi) := S_{\Lambda_m(f)}(f, \Psi), \quad \text{where } S_E(f) := S_E(f, \Psi) := \sum_{k \in E} c_k(f, \Psi)\psi_k.$$

It is clear that in the case of the normalized basis ($\|\psi_k\| = 1$, $k = 1, 2, \ldots$) the above two greedy algorithms coincide.

In the case $X = L_p$ we will write p instead of L_p in notations. It is a simple algorithm which describes the theoretical scheme for m-term approximation of an element f. We call this algorithm the Thresholding Greedy Algorithm (TGA). In order to understand the efficiency of this algorithm we compare its accuracy with the best possible when an approximant is a linear combination of m terms from Ψ. We define the best m-term approximation with regard to Ψ as follows:

$$\sigma_m(f) := \sigma_m(f, \Psi)_X := \inf_{c_k, \Lambda} \left\| f - \sum_{k \in \Lambda} c_k \psi_k \right\|_X,$$

where inf is taken over coefficients c_k and sets of indices Λ with cardinality $|\Lambda| = m$. The best we can achieve with the algorithm G_m is

$$\|f - G_m(f, \Psi, \rho)\|_X = \sigma_m(f, \Psi)_X,$$

or, a little weaker,

$$\|f - G_m(f, \Psi, \rho)\|_X \leq G\sigma_m(f, \Psi)_X \tag{2.3}$$

for all elements $f \in X$ with a constant $G = C(X, \Psi)$ independent of f and m. It is clear that in the case $X = H$ is a Hilbert space and Ψ is an orthonormal basis we have

$$\|f - G_m(f, \Psi, \rho)\|_H = \sigma_m(f, \Psi)_H.$$

Let us begin our discussion with an important class of bases: wavelet type bases. Denote $\mathcal{H} := \{H_k\}_{k=1}^{\infty}$ the Haar basis on $[0, 1)$ normalized in $L_2(0, 1)$: $H_1 = 1$ on $[0, 1)$ and for $k = 2^n + l$, $n = 0, 1, \dots$, $l = 1, 2, \dots, 2^n$,

$$H_k(x) = \begin{cases} 2^{n/2}, & x \in [(2l-2)2^{-n-1}, (2l-1)2^{-n-1}) \\ -2^{n/2}, & x \in [(2l-1)2^{-n-1}, 2l2^{-n-1}) \\ 0, & \text{otherwise.} \end{cases}$$

We denote by $\mathcal{H}_p := \{H_{k,p}\}_{k=1}^{\infty}$ the Haar basis $\mathcal{H}$ renormalized in $L_p(0, 1)$. We will use the following definition of the L_p-equivalence of bases. We say that $\Psi = \{\psi_k\}_{k=1}^{\infty}$ is L_p-equivalent to $\Phi = \{\phi_k\}_{k=1}^{\infty}$ if for any finite set Λ and any coefficients c_k, $k \in \Lambda$, we have

$$C_1(p, \Psi, \Phi)\Big\| \sum_{k\in\Lambda} c_k\phi_k \Big\|_p \le \Big\| \sum_{k\in\Lambda} c_k\psi_k \Big\|_p \le C_2(p, \Psi, \Phi)\Big\| \sum_{k\in\Lambda} c_k\phi_k \Big\|_p$$

with two positive constants $C_1(p, \Psi, \Phi), C_2(p, \Psi, \Phi)$ which may depend on p, Ψ, and Φ. For sufficient conditions on Ψ to be L_p-equivalent to $\mathcal{H}$ see [16] and [10]. In particular, it is known that all reasonable univariate wavelet type bases are L_p-equivalent to $\mathcal{H}$ for $1 < p < \infty$. We proved the following theorem in [40].

Theorem 2.1. *Let $1 < p < \infty$ and let a basis Ψ be L_p-equivalent to the Haar basis $\mathcal{H}$. Then for any $f \in L_p(0, 1)$ we have*

$$\|f - G_m^p(f, \Psi)\|_p \le C(p, \Psi)\sigma_m(f, \Psi)_p$$

with a constant $C(p, \Psi)$ independent of f and m.

By a simple renormalization argument one obtains the following version of Theorem 2.1.

Theorem 2.2. *Let $1 < p < \infty$ and let a basis Ψ be L_p-equivalent to the Haar basis $\mathcal{H}_p$. Then for any $f \in L_p(0, 1)$ and any $\rho \in D(f)$ we have*

$$\|f - G_m(f, \Psi, \rho)\|_p \le C(p, \Psi)\sigma_m(f, \Psi)_p$$

with a constant $C(p, \Psi)$ independent of f, ρ, and m.

We note that [40] also contains a generalization of Theorem 2.1 to the multivariate Haar basis obtained by the multiresolution analysis procedure. These theorems motivated us to consider the general setting of greedy approximation in Banach spaces. We concentrated on studying bases which satisfy (2.3) for all individual functions. The following Definitions 2.1, 2.2 and 2.3 are from [27].

Definition 2.1. We call a basis Ψ a *greedy basis* if for every $f \in X$ there exists a permutation $\rho \in D(f)$ such that

$$\|f - G_m(f, \Psi, \rho)\|_X \le G\sigma_m(f, \Psi)_X \tag{2.4}$$

holds with a constant independent of f, m.

The following proposition has been proved in [27].

Proposition 2.1. *If Ψ is a greedy basis, then inequality* (2.4) *holds for any permutation $\rho \in D(f)$.*

Theorem 2.2 shows that each basis Ψ which is L_p-equivalent to the univariate Haar basis $\mathcal{H}_p$ is a greedy basis for $L_p(0,1)$, $1 < p < \infty$. We note that in the case of Hilbert space each orthonormal basis is a greedy basis with a constant $G = 1$ (see (2.4)).

We give now the definitions of unconditional and democratic bases.

Definition 2.2. A basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ of a Banach space X is said to be *unconditional* if for every choice of signs $\theta = \{\theta_k\}_{k=1}^{\infty}$, $\theta_k = 1$ or -1, $k = 1, 2, \dots$, the linear operator M_θ defined by $M_\theta\big(\sum_{k=1}^{\infty} a_k \psi_k\big) = \sum_{k=1}^{\infty} a_k \theta_k \psi_k$ is a bounded operator from X into X.

Definition 2.3. We say that a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ is a *democratic basis* for X if there exists a constant $D := D(X, \Psi)$ such that for any two finite sets of indices P and Q with the same cardinality $|P| = |Q|$ we have $\big\| \sum_{k\in P} \psi_k \big\| \le D \big\| \sum_{k\in Q} \psi_k \big\|$.

We proved in [27] the following theorem.

Theorem 2.3. *A basis is greedy if and only if it is unconditional and democratic.*

This theorem gives a characterization of greedy bases. Further investigations ([41], [6], [25], [18], [21]) showed that the concept of greedy bases is very useful in direct and inverse theorems of nonlinear approximation and also in applications in statistics. The papers [27], [40] contain other results on greedy bases.

Let us discuss a question of weakening the property of a basis of being a greedy basis. We begin with the concept of quasi-greedy basis introduced in [27].

Definition 2.4. We call a basis Ψ a *quasi-greedy basis* if for every $f \in X$ and every permutation $\rho \in D(f)$ we have

$$\|G_m(f, \Psi, \rho)\|_X \le C \|f\|_X \tag{2.5}$$

with a constant C independent of f, m, and ρ.

It is clear that (2.5) is weaker then (2.4). P. Wojtaszczyk [53] proved the following theorem.

Theorem 2.4. *A basis Ψ is quasi-greedy if and only if for any $f \in X$ and any $\rho \in D(f)$ we have*

$$\|f - G_m(f, \Psi, \rho)\| \to 0 \quad \text{as } m \to \infty. \tag{2.6}$$

We proceed to an intermediate concept of almost greedy basis. This concept has been introduced and studied in [14]. Let

$$f = \sum_{k=1}^{\infty} c_k(f)\psi_k.$$

We define the following expansional best m-term approximation of f:

$$\tilde{\sigma}_m(f) := \tilde{\sigma}_m(f, \Psi) := \inf_{\Lambda, |\Lambda|=m} \Big\| f - \sum_{k\in\Lambda} c_k(f)\psi_k \Big\|.$$

It is clear that

$$\sigma_m(f, \Psi) \le \tilde{\sigma}_m(f, \Psi).$$

It is also clear that for an unconditional basis Ψ we have

$$\tilde{\sigma}_m(f, \Psi) \le C\sigma_m(f, \Psi).$$

Definition 2.5. We call a basis Ψ *almost greedy* if for every $f \in X$ there exists a permutation $\rho \in D(f)$ such that

$$\|f - G_m(f, \Psi, \rho)\|_X \le C\tilde{\sigma}_m(f, \Psi)_X \tag{2.7}$$

holds with a constant independent of f, m.

The following proposition follows from the proof of Theorem 3.3 of [14] (see Theorem 2.5 below).

Proposition 2.2. *If Ψ is an almost greedy basis then* (2.7) *holds for any permutation $\rho \in D(f)$.*

The following characterization of almost greedy bases has been obtained in [14].

Theorem 2.5. *Suppose Ψ is a basis of a Banach space. The following are equivalent:*

A. *Ψ is almost greedy.*

B. *Ψ is quasi-greedy and democratic.*

C. *For any (respectively, every) $\lambda > 1$ there is a constant $C = C_\lambda$ such that*

$$\|f - G_{[\lambda m]}(f, \Psi)\| \le C_\lambda \sigma_m(f, \Psi).$$

We now proceed to a generalization of the concept of greedy bases from [26] that is useful in statistical applications. Let Ψ be a basis for X. If $\inf_k \|\psi_k\| > 0$ then $c_k(f) \to 0$ as $k \to \infty$, where

$$f = \sum_{k=1}^{\infty} c_k(f)\psi_k.$$

Then we can rearrange the coefficients $\{c_k(f)\}$ in the decreasing way

$$|c_{k_1}(f)| \geq |c_{k_2}(f)| \geq \cdots$$

and define the mth greedy approximant as

$$G_m(f, \Psi) := \sum_{j=1}^{m} c_{k_j}(f)\psi_{k_j}. \tag{2.8}$$

In the case $\inf_k \|\psi_k\| = 0$ we define $G_m(f, \Psi)$ by (2.8) for f of the form

$$f = \sum_{k \in Y} c_k(f)\psi_k, \quad |Y| < \infty. \tag{2.9}$$

Let a weight sequence $w = \{w_k\}_{k=1}^{\infty}$, $w_k > 0$, be given. For $\Lambda \subset \mathbb{N}$ denote $w(\Lambda) := \sum_{k \in \Lambda} w_k$. For a positive real number $v > 0$ define

$$\sigma_v^w(f, \Psi) := \inf_{\{b_k\}, \Lambda : w(\Lambda) \leq v} \|f - \sum_{k \in \Lambda} b_k \psi_k\|,$$

where the sets Λ are finite.

Definition 2.6. We call a basis Ψ a *weight-greedy basis* (w-greedy basis) if for any $f \in X$ in the case $\inf_k \|\psi_k\| > 0$ or for any $f \in X$ of the form (2.9) in the case $\inf_k \|\psi_k\| = 0$ we have

$$\|f - G_m(f, \Psi)\| \leq C_G \sigma^w_{w(\Lambda_m)}(f, \Psi),$$

where

$$G_m(f, \Psi) = \sum_{k \in \Lambda_m} c_k(f)\psi_k, \quad |\Lambda_m| = m.$$

Definition 2.7. We call a basis Ψ *weight-democratic* (w-democratic basis) if for any finite $A, B \subset \mathbb{N}$ such that $w(A) \leq w(B)$ we have

$$\|\sum_{k \in A} \psi_k\| \leq C_D \|\sum_{k \in B} \psi_k\|.$$

Recently, we proved in [26] the following criterion for w-greedy bases.

Theorem 2.6. *A basis Ψ is a w-greedy basis if and only if it is unconditional and w-democratic.*

The reader can find a further discussion in the surveys [8], [28], [46], [47].

3. Optimal methods in nonlinear approximation

In the widths problem of linear approximation (see Section 1) we were looking for an optimal n-dimensional subspace for approximating a given function class. A nonlinear analogue of this setting is the following. Let a function class F and a Banach space X be given. Assume that on the basis of some additional information we know that our basis for m-term approximation should satisfy some structural properties, for instance, it has to be orthogonal. Then similarly to the setting for the widths d_n, λ_n, φ_n we get the optimization problems for m-term nonlinear approximation. Let $\mathbb{B}$ be a collection of bases satisfying a given property.

I. Define an analogue of the Kolmogorov width

$$\sigma_m(F, \mathbb{B})_X := \inf_{\Psi \in \mathbb{B}} \sup_{f \in F} \sigma_m(f, \Psi)_X.$$

II. Define an analogue of the orthowidth

$$\gamma_m(F, \mathbb{B})_X := \inf_{\Psi \in \mathbb{B}} \sup_{f \in F} \|f - G_m(f, \Psi)\|_X.$$

We present here some results in the case $\mathbb{B} = \mathbb{O}$, the set of orthonormal bases, $F = W_q^r$, $X = L_p$, $1 \le q, p \le \infty$. First of all we formulate a result (see [24], [43]) that shows that in the case $p < 2$ we need some more restrictions on $\mathbb{B}$ in order to obtain meaningful results (lower bounds).

Proposition 3.1. *For any $1 \le p < 2$ there exists a complete in $L_2(0, 1)$ orthonormal system Φ such that for each $f \in L_p(0, 1)$ we have $\sigma_1(f, \Phi)_p = 0$.*

Let us restrict our further discussion to the case $p \ge 2$. This case was also more interesting in the linear approximation discussion (see Section 1). Kashin [23] proved that

$$\sigma_m(W_\infty^r, \mathbb{O})_2 \gg m^{-r}. \tag{3.1}$$

We proved (see [11]) that

$$\sigma_m(W_2^r, \mathcal{T})_\infty \ll m^{-r}. \tag{3.2}$$

The estimates (3.1) and (3.2) imply that for $2 \le q, p \le \infty$ we have

$$\sigma_m(W_q^r, \mathbb{O})_p \asymp \sigma_m(W_q^r, \mathcal{T})_p \asymp m^{-r}. \tag{3.3}$$

Let us compare this relation with (1.2). We see that the best m-term trigonometric approximation provides the same accuracy as the best approximation from an optimal m-dimensional subspace. An advantage of nonlinear approximation here is that we use a natural basis instead of an existing but nonconstructive subspace. However, we should note that the estimate (3.2) was proved in [11] as an existence theorem. We did not give an algorithm to get (3.2) in [11]. We gave such an algorithm in [49]

(see a further discussion in Section 6). The Thresholding Greedy Algorithm does not provide the estimate (3.2). We have (see [42])

$$\sup_{f\in W_2^r}\|f-G_m(f,\mathcal{T})\|_\infty \asymp m^{-r+1/2}.$$

It is known from different results (see [9], [8], [45]) that wavelets are well designed for nonlinear approximation.

In the multivariate periodic case the following basis $U^d := U\times\cdots\times U$ has approximation properties close to the corresponding properties of wavelets. We define the system $U := \{U_I\}$ in the univariate case. Denote

$$\begin{aligned}
U_n^+(x) &:= \sum_{k=0}^{2^n-1} e^{ikx} = \frac{e^{i2^nx}-1}{e^{ix}-1}, \qquad n=0,1,2,\dots;\\
U_{n,k}^+(x) &:= e^{i2^nx}U_n^+(x-2\pi k2^{-n}), \qquad k=0,1,\dots,2^n-1;\\
U_{n,k}^-(x) &:= e^{-i2^nx}U_n^+(-x+2\pi k2^{-n}), \quad k=0,1,\dots,2^n-1.
\end{aligned}$$

We normalize the system of functions $\{U_{n,k}^+, U_{n,k}^-\}$ in L_2 and enumerate it by dyadic intervals. We write

$$\begin{aligned}
U_I(x) &:= 2^{-n/2}U_{n,k}^+(x) \quad \text{with } I=[(k+1/2)2^{-n},(k+1)2^{-n});\\
U_I(x) &:= 2^{-n/2}U_{n,k}^-(x) \quad \text{with } I=[k2^{-n},(k+1/2)2^{-n});
\end{aligned}$$

and

$$U_{[0,1)}(x) := 1.$$

P. Wojtaszczyk [52] proved that the system U is an unconditional basis for $L_p(\mathbb{T})$, $1<p<\infty$.

We define the anisotropic multivariate periodic Hölder–Nikol'skii classes NH_p^R in the following way. The class NH_p^R, $R=(R_1,\dots,R_d)$ and $1\le p\le\infty$, is the set of periodic functions $f\in L_p([0,2\pi]^d)$ such that for each $l_j=[R_j]+1$, $j=1,\dots,d$, the following relations hold:

$$\|f\|_p\le 1, \quad \|\Delta_t^{l_j,j}f\|_p\le |t|^{R_j}, \quad j=1,\dots,d, \tag{3.4}$$

where $\Delta_t^{l,j}$ is the l-th difference with step t in the variable x_j. In the case $d=1$ NH_p^R coincides with the standard Hölder class H_p^R. For $R=(R_1,\dots,R_d)$, $R_j>0$, $j=1,\dots,d$, we define $g(R) := \left(\sum_{j=1}^d R_j^{-1}\right)^{-1}$. The following result has been proved in [45].

Theorem 3.1. *Let $1<q,p<\infty$. Then for R such that $g(R)>(1/q-1/p)_+$ we have*

$$\sup_{f\in NH_q^R}\|f-G_m^{L_p}(f,U^d)\|_p \ll m^{-g(R)}.$$

We also proved in [45] that the basis U^d is an optimal orthonormal basis for approximation of classes NH_q^R in L_p:

$$\sigma_m(NH_q^R, \mathbb{O})_p \asymp \sigma_m(NH_q^R, U^d)_p \asymp m^{-g(R)} \tag{3.5}$$

for $1 < q < \infty$, $2 \le p < \infty$, $g(R) > (1/q - 1/p)_+$. It is important to remark that Theorem 3.1 guarantees that the estimate in (3.5) can be realized by the TGA with regard to U^d.

4. The TGA with regard to the trigonometric system

Let us consider nonlinear approximation with regard to the trigonometric system $\mathcal{T}^d := \mathcal{T} \times \cdots \times \mathcal{T}$ (d times). The existence of best m-term trigonometric approximation was proved in [2] (see also [42]). The method $G_m(f) := G_m(f, \mathcal{T}^d)$ has an advantage over the traditional approximation by trigonometric polynomials in the case of approximation of functions of several variables. In this case ($d > 1$) there is no natural order of trigonometric system and the use of G_m allows us to avoid the problem of finding natural subspaces of trigonometric polynomials for approximation purposes. We proved in [42] the following results.

Theorem 4.1. *For each $f \in L_p(\mathbb{T}^d)$ we have*

$$\|f - G_m(f)\|_p \le (1 + 3m^{h(p)})\sigma_m(f)_p, \quad 1 \le p \le \infty,$$

where $h(p) := |1/2 - 1/p|$.

Remark 4.1. For all $1 \le p \le \infty$

$$\|G_m(f)\|_p \le m^{h(p)}\|f\|_p.$$

Remark 4.2. There is a positive absolute constant C such that for each m and $1 \le p \le \infty$ there exists a function $f \ne 0$ with the property

$$\|G_m(f)\|_p \ge Cm^{h(p)}\|f\|_p. \tag{4.1}$$

The above results show that the trigonometric system is not a quasi-greedy basis for L_p, $p \ne 2$. This leads to a natural attempt to consider some other algorithms that may have some advantages over the TGA in the case of $\mathcal{T}$. We discuss here the performance of the WCGA (see Section 6) with regard to $\mathcal{T}$.

Let us compare the rate of approximation of the TGA and the WCGA. Let $\mathcal{RT}$ denote the real trigonometric system $1/2, \sin x, \cos x, \dots$. We need to switch to this system from the complex trigonometric system because the algorithm WCGA is defined for the real Banach space. We note that the system $\mathcal{RT}$ is not normalized in L_p but quasinormalized: $C_1 \le \|t\|_p \le C_2$ for any $t \in \mathcal{RT}$ with absolute constants

C_1, C_2, $1 \le p \le \infty$. It is sufficient for the application of the general methods developed in Section 6. For a function f with absolutely convergent Fourier series

$$f(x) = a_0/2 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx)$$

denote

$$\|f\|_A := |a_0| + \sum_{k=1}^{\infty}(|a_k| + |b_k|).$$

Define the class

$$A := A(\mathcal{RT}) := \{f : \|f\|_A \le 1\}.$$

For a sequence $\tau := \{t_k\}$ with $t_k = t$, $k = 1, 2, \dots$, we replace τ by t in the notation. Theorem 6.1 and (6.2) imply the following result.

Theorem 4.2. *Let $0 < t \le 1$. For $f \in A$ we have*

$$\|f - G_m^{c,t}(f, \mathcal{RT})\|_p \le C(p,t)m^{-1/2}, \quad 2 \le p < \infty. \tag{4.2}$$

This estimate and Theorem 4.1 imply that for $f \in A$ we have

$$\|f - G_m(f, \mathcal{RT})\|_p \le C(p,t)m^{-1/p}, \quad 2 \le p < \infty, \tag{4.3}$$

which is weaker than (4.2). It is proved in [15] that (4.3) can not be improved. Thus the WCGA works better than the TGA for the class A. We note that the restriction $p < \infty$ in (4.2) is important. We gave a lower estimate for m-term approximation in L_∞ in [47].

Proposition 4.1. *For a given m define*

$$f := \sum_{k=0}^{2m} \cos 3^k x.$$

Then we have

$$\sigma_m(f, \mathcal{T})_\infty \ge m/8.$$

5. Convergence of the TGA with regard to the trigonometric system

We discuss in this section the following nonlinear method of summation of trigonometric Fourier series. Consider a periodic function $f \in L_p(\mathbb{T}^d)$, $1 \le p \le \infty$, $(L_\infty(\mathbb{T}^d) = C(\mathbb{T}^d))$, defined on the d-dimensional torus $\mathbb{T}^d$. Take a number $t \in (0, 1]$. Let a number $m \in \mathbb{N}$ be given and Λ_m be a set of $k \in \mathbb{Z}^d$ with the properties

$$\min_{k \in \Lambda_m} |\hat{f}(k)| \ge t \max_{k \notin \Lambda_m} |\hat{f}(k)|, \quad |\Lambda_m| = m, \tag{5.1}$$

where

$$\hat{f}(k) := (2\pi)^{-d} \int_{\mathbb{T}^d} f(x) e^{-i(k,x)} dx$$

is a Fourier coefficient of f. We define

$$G_m^t(f) := G_m^t(f, \mathcal{T}^d) := S_{\Lambda_m}(f) := \sum_{k \in \Lambda_m} \hat{f}(k) e^{i(k,x)}$$

and call it an mth weak greedy approximant of f with regard to the trigonometric system $\mathcal{T}^d = \{e^{i(k,x)}\}_{k \in \mathbb{Z}^d}$. We write $G_m(f) = G_m^1(f)$ and call it an mth greedy approximant. Clearly, an mth weak greedy approximant and even an mth greedy approximant may not be unique. Here we do not impose any extra restrictions on Λ_m in addition to (5.1). Thus theorems formulated below hold for any choice of Λ_m satisfying (5.1) or, in other words, for any realization $G_m^t(f)$ of the weak greedy approximation.

There has recently been (see the surveys [8], [47], [28]) much interest in approximation of functions by m-term approximants with regard to a basis (or minimal system). In this section we will discuss in detail only results concerning the trigonometric system. Answering a question raised by Carleson and Coifman, T. W. Körner constructed in [31] a function from $L_2(\mathbb{T})$ and then in [32] a continuous function such that $\{G_m(f, \mathcal{T})\}$ diverges almost everywhere. It has been proved in [42] for $p \neq 2$ and in [7] for $p < 2$ that there exists an $f \in L_p(\mathbb{T})$ such that $\{G_m(f, \mathcal{T})\}$ does not converge in L_p. It was remarked in [47] that the method from [42] gives a little bit more: 1) There exists a continuous function f such that $\{G_m(f, \mathcal{T})\}$ does not converge in $L_p(\mathbb{T})$ for any $p > 2$; and 2) there exists a function f that belongs to any $L_p(\mathbb{T})$, $p < 2$, such that $\{G_m(f, \mathcal{T})\}$ does not converge in measure. Thus the above negative results show that the condition $f \in L_p(\mathbb{T}^d)$, $p \neq 2$, does not guarantee convergence of $\{G_m(f, \mathcal{T}^d)\}$ in the L_p-norm. The main goal of this section is to discuss an additional (to $f \in L_p$) condition on f to guarantee that $\|f - G_m(f, \mathcal{T}^d)\|_p \to 0$ as $m \to \infty$. Some results in this direction have been obtained in [29], [30]. In the case $2 < p \leq \infty$ we found in [29] necessary and sufficient conditions on a decreasing sequence $\{A_n\}_{n=1}^{\infty}$ to guarantee the L_p-convergence of $\{G_m(f)\}$ for all $f \in L_p$, satisfying $a_n(f) \leq A_n$, where $\{a_n(f)\}$ is a decreasing rearrangement of absolute values of the Fourier coefficients of f. We will formulate three theorems from [29].

For $f \in L_1(\mathbb{T}^d)$ let $\{\hat{f}(k(l))\}_{l=1}^{\infty}$ denote the decreasing rearrangement of $\{\hat{f}(k)\}_{k \in \mathbb{Z}^d}$, i.e.

$$|\hat{f}(k(1))| \geq |\hat{f}(k(2))| \geq \cdots . \tag{5.2}$$

Denote $a_n(f) := |\hat{f}(k(n))|$.

Theorem 5.1. *Let $2 < p < \infty$ and let a decreasing sequence $\{A_n\}_{n=1}^{\infty}$ satisfy the condition*

$$A_n = o(n^{1/p-1}) \quad \text{as } n \to \infty. \tag{5.3}$$

Then for any $f \in L_p(\mathbb{T}^d)$ *with the property* $a_n(f) \le A_n$, $n = 1, 2, \dots$, *we have*

$$\lim_{m\to\infty} \|f - G_m^t(f, \mathcal{T})\|_p = 0. \tag{5.4}$$

We also proved in [29] that for any decreasing sequence $\{A_n\}$ satisfying

$$\limsup_{n\to\infty} A_n n^{1-1/p} > 0$$

there exists a function $f \in L_p$ such that $a_n(f) \le A_n$, $n = 1, \dots$, with divergent in the L_p norm sequence of greedy approximants $\{G_m(f)\}$.

Theorem 5.2. *Let a decreasing sequence* $\{A_n\}_{n=1}^{\infty}$ *satisfy the following condition* $(\mathcal{A}_\infty)$*:*

$$\sum_{M<n\le e^M} A_n = o(1) \quad \text{as } M \to \infty. \tag{5.5}$$

Then for any $f \in C(\mathbb{T})$ *with the property* $a_n(f) \le A_n$, $n = 1, 2, \dots$, *we have*

$$\lim_{m\to\infty} \|f - G_m^t(f, \mathcal{T})\|_\infty = 0. \tag{5.6}$$

The following theorem from [29] shows that the condition $(\mathcal{A}_\infty)$ in Theorem 5.2 is sharp.

Theorem 5.3. *Assume that a decreasing sequence* $\{A_n\}_{n=1}^{\infty}$ *does not satisfy the condition* $(\mathcal{A}_\infty)$. *Then there exists a function* $f \in C(\mathbb{T})$ *with the property* $a_n(f) \le A_n$, $n = 1, 2, \dots$, *and such that we have*

$$\limsup_{m\to\infty} \|f - G_m(f, \mathcal{T})\|_\infty > 0$$

for some realization $G_m(f, \mathcal{T})$.

In [30] we concentrated on imposing extra conditions in the following form. We assume that for some sequence $\{M(m)\}$, $M(m) > m$, we have

$$\|G_{M(m)}(f) - G_m(f)\|_p \to 0 \quad \text{as } m \to \infty. \tag{5.7}$$

In the case that p is an even number or $p = \infty$ we found in [30] necessary and sufficient conditions on the growth of the sequence $\{M(m)\}$ to provide convergence $\|f - G_m(f)\|_p \to 0$ as $m \to \infty$. We proved the next theorem in [30].

Theorem 5.4. *Let* $p = 2q$, $q \in \mathbb{N}$, *be an even integer,* $\delta > 0$. *Assume that* $f \in L_p(\mathbb{T})$ *and there exists a sequence of positive integers* $M(m) > m^{1+\delta}$ *such that*

$$\|G_m(f) - G_{M(m)}(f)\|_p \to 0 \quad \text{as } m \to \infty.$$

Then we have

$$\|G_m(f) - f\|_p \to 0 \quad \text{as } m \to \infty.$$

In [30] we proved that the condition $M(m) > m^{1+\delta}$ cannot be replaced by the condition $M(m) > m^{1+o(1)}$.

Theorem 5.5. *For any $p \in (2, \infty)$ there exists a function $f \in L_p(\mathbb{T})$ with divergent in the $L_p(\mathbb{T})$ norm sequence $\{G_m(f)\}$ of greedy approximations with the following property. For any sequence $\{M(m)\}$ such that $m \le M(m) \le m^{1+o(1)}$ we have*

$$\|G_{M(m)}(f) - G_m(f)\|_p \to 0 \quad (m \to 0).$$

In [30] we also considered the case $p = \infty$. We proved there necessary and sufficient conditions for convergence of greedy approximations in the uniform norm. For a mapping $\alpha : W \to W$ we denote by α_k its k-fold iteration: $\alpha_k := \alpha \circ \alpha_{k-1}$.

Theorem 5.6. *Let $\alpha : \mathbb{N} \to \mathbb{N}$ be strictly increasing. Then the following conditions are equivalent:*

a) *For some $k \in \mathbb{N}$ and for any sufficiently large $m \in \mathbb{N}$ we have $\alpha_k(m) > e^m$.*

b) *If $f \in C(\mathbb{T})$ and*

$$\left\|G_{\alpha(m)}(f) - G_m(f)\right\|_\infty \to 0 \quad (m \to \infty)$$

then

$$\|f - G_m(f)\|_\infty \to 0 \quad (m \to \infty).$$

The proof of the necessary condition is based on the above Theorem 5.3 from [29]. In the proof of the sufficient condition we use the following special inequality (see [30]).

By $\Sigma_m(\mathcal{T})$ we denote the set of all trigonometric polynomials with at most m nonzero coefficients.

Theorem 5.7. *For any $h \in \Sigma_m(\mathcal{T})$ and any $g \in L_\infty$ one has*

$$\|h + g\|_\infty \ge K^{-2}\|h\|_\infty - e^{C(K)m}\|\{\hat{g}(k)\}\|_{\ell_\infty}, \quad K > 1. \tag{5.8}$$

We note that in the proof of the above inequality we used a deep result on the uniform approximation property of the space $C(X)$ (see [4]). The paper [30] contains some other inequalities in the style of (5.8).

6. General greedy algorithms

The purpose of this section is to discuss nonlinear m-term approximation and greedy algorithms with regard to a general system (dictionary). We concentrate here on a discussion of m-term approximation with regard to redundant dictionaries in Banach spaces. We will discuss only one example of an algorithm from the family of greedy

algorithms. The reader can find a further discussion of greedy approximation in Banach spaces in the survey [47]. This section is based on the paper [44] which in turn is a combination of ideas and methods developed for Banach spaces in a fundamental paper [13] with the approach used in [50] in the case of Hilbert spaces. The papers [13] and [50] contain detailed historical remarks and we refer the reader to those papers. Two greedy type approximation methods the Weak Chebyshev Greedy Algorithm (WCGA) and the Weak Relaxed Greedy Algorithm (WRGA) have been introduced and studied in [44]. These methods (WCGA and WRGA) are very general approximation methods that work well in an arbitrary uniformly smooth Banach space X for any dictionary $\mathcal{D}$ (see below). Surprisingly, it turned out that these general approximation methods are also very good for specific dictionaries. It has been observed in [15] that the WCGA provides constructive methods in m-term trigonometric approximation in L_p, $p \in [2, \infty)$, which realizes optimal rate of m-term approximation for different function classes. In [48] the WCGA and WRGA have been used in constructing deterministic cubature formulas for a wide variety of function classes with error estimates similar to those for the Monte Carlo Method. It looks like WCGA and WRGA can be considered as a constructive deterministic alternative to (substitute for) some powerful probabilistic methods. This observation encouraged us to continue thorough study of WCGA and WRGA.

In this section we discuss in detail only WCGA. In [44] we developed the theory of the Weak Chebyshev Greedy Algorithm in a general setting: X is an arbitrary uniformly smooth Banach space and $\mathcal{D}$ is any dictionary. We keep the term *greedy algorithm* in the name of this approximation method for two reasons. First, this term has been used in previous papers and has become a standard name for procedures like WCGA. For more discussion of the terminology see [47, Remark 1.1, p. 38]. Second, clearly, in the above general setting the term *algorithm* cannot be confused with the same term used in a more restricted sense, say, in computer science. We note that in the case of finite dimensional X and finite $\mathcal{D}$ the above methods are algorithms in a strict sense.

In this section we discuss the following two applications of general greedy algorithms from [49]. In [49] we used WCGA to build a constructive method for m-term trigonometric approximation in the uniform norm. It is known that the case of approximating by m-term trigonometric polynomials in the uniform norm is the most difficult. We note that in the case of L_p-norms with $p < \infty$ the corresponding constructive method has been provided in [15]. In [49] we also studied a slight modification of incremental type algorithm from [13]. We applied that algorithm for constructing deterministic sets of points with small L_p discrepancy and also with small symmetrized L_p discrepancy.

We now proceed to a systematic presentation of the mentioned above results. Let X be a Banach space with norm $\|\cdot\|$. We say that a set of elements (functions) $\mathcal{D}$ from X is a dictionary if each $g \in \mathcal{D}$ has norm less than or equal to one ($\|g\| \le 1$),

$$g \in \mathcal{D} \quad \text{implies} \quad -g \in \mathcal{D},$$

and $\overline{\operatorname{span}}\mathcal{D} = X$. We note that in [44] we required in the definition of a dictionary normalization of its elements ($\|g\| = 1$). However, it is pointed out in [49] that it is easy to check that the arguments from [44] work under the assumption $\|g\| \le 1$ instead of $\|g\| = 1$. In applications it is more convenient for us to have the assumption $\|g\| \le 1$ than normalization of a dictionary.

For an element $f \in X$ we denote by F_f a norming (peak) functional for f:

$$\|F_f\| = 1, \quad F_f(f) = \|f\|.$$

The existence of such a functional is guaranteed by the Hahn–Banach theorem. Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given sequence of nonnegative numbers $t_k \le 1$, $k = 1, \dots$. We define (see [44]) the Weak Chebyshev Greedy Algorithm (WCGA) which is a generalization for Banach spaces of the Weak Orthogonal Greedy Algorithm defined and studied in [50] (see also [12] for the Orthogonal Greedy Algorithm).

6.1. Weak Chebyshev Greedy Algorithm (WCGA). We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \ge 1$ we inductively define

1) $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}^c}(\varphi_m^c) \ge t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}^c}(g).$$

2) Define

$$\Phi_m := \Phi_m^{\tau} := \operatorname{span}\{\varphi_j^c\}_{j=1}^m,$$

and define $G_m^c := G_m^{c,\tau}$ to be the best approximant to f from Φ_m.

3) Denote

$$f_m^c := f_m^{c,\tau} := f - G_m^c.$$

The term "weak" in this definition means that at step 1) we do not shoot for the optimal element of the dictionary which realizes the corresponding supremum but are satisfied with the weaker property rather than being optimal. The obvious reason for this is that we do not know in general that the optimal one exists. Another practical reason is that, the weaker the assumption, the easier to satisfy it and, therefore, easier to realize in practice.

We consider here approximation in uniformly smooth Banach spaces. For a Banach space X we define the modulus of smoothness by

$$\rho(u) := \sup_{\|x\|=\|y\|=1} \left(\frac{1}{2}(\|x + uy\| + \|x - uy\|) - 1 \right).$$

The uniformly smooth Banach space is the one with the property

$$\lim_{u \to 0} \rho(u)/u = 0.$$

It is easy to see that for any Banach space X its modulus of smoothness $\rho(u)$ is an even convex function satisfying the inequalities

$$\max(0, u-1) \le \rho(u) \le u, \quad u \in (0, \infty). \tag{6.1}$$

It is well known (see for instance [13], Lemma B.1) that in the case $X = L_p$, $1 \le p < \infty$ we have

$$\rho(u) \le \begin{cases} u^p/p & \text{if} \quad 1 \le p \le 2, \\ (p-1)u^2/2 & \text{if} \quad 2 \le p < \infty. \end{cases} \tag{6.2}$$

It is also known (see [34], p. 63) that for any X with $\dim X = \infty$ one has

$$\rho(u) \ge (1+u^2)^{1/2} - 1$$

and for every X, $\dim X \ge 2$,

$$\rho(u) \ge Cu^2, \quad C > 0.$$

This limits power type moduli of smoothness of nontrivial Banach spaces to the case $1 \le q \le 2$. Denote by $A(\mathcal{D})$ the closure of the convex hull of $\mathcal{D}$. The following theorem from [44] gives the rate of convergence of the WCGA for f in $A(\mathcal{D})$.

Theorem 6.1. *Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Then for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \le 1$, $k = 1, 2, \dots$, we have for any $f \in A(\mathcal{D})$ that*

$$\|f - G_m^{c,\tau}(f, \mathcal{D})\| \le C(q, \gamma)\Big(1 + \sum_{k=1}^{m} t_k^p\Big)^{-1/p}, \quad p := \frac{q}{q-1},$$

with a constant $C(q, \gamma)$ which may depend only on q and γ.

In [49] we demonstrated the power of the WCGA in classical areas of harmonic analysis. The problem concerns the trigonometric m-term approximation in the uniform norm. Let $\mathcal{RT}(N)$ be the subspace of real trigonometric polynomials of order N. Both R. S. Ismagilov [20] and V. E. Maiorov [35] used constructive methods to get their estimates (1.6) and (1.8). V. E. Maiorov [35] applied a number theoretical method based on Gaussian sums. The key point of that technique can be formulated in terms of best m-term approximation of trigonometric polynomials. Using the Gaussian sums one can prove (constructively) the estimate

$$\sigma_m(t, \mathcal{RT})_\infty \le CN^{3/2}m^{-1}\|t\|_1, \quad t \in \mathcal{RT}(N). \tag{6.6}$$

Denote

$$\Big\|a_0/2 + \sum_{k=1}^{N}(a_k \cos kx + b_k \sin kx)\Big\|_A := |a_0| + \sum_{k=1}^{N}(|a_k| + |b_k|).$$

We note that by the simple inequality

$$\|t\|_A \le (2N+1)\|t\|_1, \quad t \in \mathcal{RT}(N),$$

the estimate (6.6) follows from the estimate

$$\sigma_m(t, \mathcal{RT})_\infty \le C(N^{1/2}/m)\|t\|_A, \quad t \in \mathcal{RT}(N). \tag{6.7}$$

Thus (6.7) is stronger than (6.6). The following estimate was proved in [11]:

$$\sigma_m(t, \mathcal{RT})_\infty \le Cm^{-1/2}(\ln(1+N/m))^{1/2}\|t\|_A, \quad t \in \mathcal{RT}(N). \tag{6.8}$$

In a way (6.8) is much stronger than (6.7) and (6.6). The proof of (6.8) from [11] is not constructive. The estimate (6.8) has been proved in [11] with the help of a nonconstructive theorem of Gluskin [17]. In [49] we gave a constructive proof of (6.8). The key ingredient of that proof is the WCGA. In the paper [15] we already pointed out that the WCGA provides a constructive proof of the estimate

$$\sigma_m(f, \mathcal{T})_p \le C(p)m^{-1/2}\|f\|_A, \quad p \in [2, \infty). \tag{6.9}$$

The known proofs (before [15]) of (6.9) were nonconstructive (see discussion in [15, Section 5]).

We formulate here a result from [49] (see Theorem 4.1).

Theorem 6.2. *There exists a constructive method $A(N, m)$ such that for any $t \in \mathcal{RT}(N)$ it provides an m-term trigonometric polynomial $A(N, m)(t)$ with the following approximation property:*

$$\|t - A(N, m)(t)\|_\infty \le Cm^{-1/2}(\ln(1+N/m))^{1/2}\|t\|_A$$

with an absolute constant C.

In [49] we applied greedy type algorithms for constructing points with small discrepancy and small symmetrized discrepancy. Let $1 \le p \le \infty$. We will define first the L_p discrepancy (the L_p-star discrepancy) of points $\{\xi^1, \dots, \xi^m\} \subset \Omega_d := [0,1]^d$. Let $\chi_{[a,b]}(\cdot)$ be a characteristic function of the interval $[a, b]$. Denote for $x, y \in \Omega_d$

$$B(x, y) := \prod_{j=1}^d \chi_{[0,x_j]}(y_j).$$

Then the L_p discrepancy of $\xi := \{\xi^1, \dots, \xi^m\} \subset \Omega_d$ is defined by

$$D(\xi, m, d)_p := \Big\| \int_{\Omega_d} B(x, y)dy - \frac{1}{m}\sum_{\mu=1}^m B(x, \xi^\mu) \Big\|_{L_p(\Omega_d)}.$$

We are interested in ξ with small discrepancy. Consider

$$D(m, d)_p := \inf_\xi D(\xi, m, d)_p.$$

The concept of discrepancy is a fundamental concept in numerical integration. There are many books and survey papers on discrepancy and related topics. We mention some of them as a reference for the history of the subject: [33], [3], [36], [5], [37], [48]. For $1 < p < \infty$ the following relation is known (see [3, p. 5]):

$$D(m,d)_p \asymp m^{-1}(\ln m)^{(d-1)/2} \tag{6.10}$$

with constants in $\asymp$ depending on p and d. The right order of $D(m,d)_p$, $p = 1, \infty$, for $d \geq 3$ is unknown. Recently, driven by possible applications (see [37]) in numerical integration the tendancy to control dependence of $D(m,d)_p$ on both variables m and d has appeared. Very interesting results in this direction have been obtained in [19]. The authors established the estimate

$$D(m,d)_\infty \leq Cd^{1/2}m^{-1/2} \tag{6.11}$$

with C an absolute constant. It is pointed out in [19] that (6.11) is only an existence theorem and even a constant C in (6.11) is unknown. The proof is a probabilistic one. There are also some other estimates in [19] with explicit constants. We mention one of them:

$$D(m,d)_\infty \leq C(d\ln d)^{1/2}((\ln m)/m)^{1/2} \tag{6.12}$$

with an explicit constant C. The proof of (6.12) is also probabilistic.

In [49] we gave constructive proofs of the following two upper estimates:

$$D(m,d)_p \leq C_1 p^{1/2}m^{-1/2}, \qquad p \in [2,\infty),$$

$$D(m,d)_\infty \leq C_2 d^{3/2}(\max(\ln d, \ln m))^{1/2}m^{-1/2}, \quad d,m \geq 2,$$

with effective absolute constants C_1 and C_2. The term *constructive proof* goes back to Kronecker who outlined the program of giving constructive proofs of theorems that were established as existence theorems. Following traditions of approximation theory we understand constructive proof as a proof that provides a construction of an object and this construction has a potential of being implemented numerically. For instance, a proof by contradiction or a probabilistic proof establishing existence of an object is not a constructive proof for us. In [49] we provided a method which consists of maximizing (approximately) certain functions of d variables at each step. For a given $p \in [2,\infty)$ after m steps of this method we obtain a set $\xi = \{\xi^1, \dots, \xi^m\} \subset \Omega_d$ of points with small L_p discrepancy

$$D(\xi, m, d)_p \leq C_1 p^{1/2}m^{-1/2}$$

with effective absolute constant C_1. The above method is a greedy type algorithm (see the IA(ϵ) below) which is a slight modification of the corresponding procedure from [13]. Here we do not assume that a dictionary $\mathcal{D}$ is symmetric: $g \in \mathcal{D}$ implies $-g \in \mathcal{D}$. To indicate this we will use the notation $\mathcal{D}^+$ for such a dictionary. We do not assume that elements of a dictionary $\mathcal{D}^+$ are normalized ($\|g\| = 1$ if $g \in \mathcal{D}^+$) we only assume that $\|g\| \leq 1$ if $g \in \mathcal{D}^+$. By $\mathcal{A}_1(\mathcal{D}^+)$ we denote the closure of the convex hull of $\mathcal{D}^+$. Let $\epsilon = \{\epsilon_n\}_{n=1}^\infty$, $\epsilon_n > 0$, $n = 1, 2, \dots$.

6.2. Incremental algorithm with schedule ϵ (IA(ϵ)). Let $f \in \mathcal{A}_1(\mathcal{D}^+)$. Denote $f_0^{i,\epsilon} := f$ and $G_0^{i,\epsilon} := 0$. Then for each $m \geq 1$ we inductively define

1. $\varphi_m^{i,\epsilon} \in \mathcal{D}^+$ is any element satisfying

$$F_{f_{m-1}^{i,\epsilon}}(\varphi_m^{i,\epsilon} - f) \geq -\epsilon_m.$$

2. Define

$$G_m^{i,\epsilon} := (1 - 1/m)G_{m-1}^{i,\epsilon} + \varphi_m^{i,\epsilon}/m.$$

3. Denote

$$f_m^{i,\epsilon} := f - G_m^{i,\epsilon}.$$

References

[1] Andrianov, A. V., Temlyakov, V. N., Best m-term approximation of functions from classes MW_q^r. In *Approximation Theory IX*, Innov. Appl. Math., Vanderbilt University Press, Nashville, TN, 1998, 7–14.

[2] Baishanski, B. M., Approximation by polynomials of given length. *Illinois J. Math.* **27** (1983), 449–458.

[3] Beck, J., Chen, W., *Irregularities of distribution*. Cambridge Tracts in Math. 89, Cambridge University Press, Cambridge 1987.

[4] Bourgain, J., A remark on the behaviour of L^p-multipliers and the range of operators acting on L^p-spaces. *Israel J. Math.* **79** (1992), 193–206.

[5] Chazelle, B., *The Discrepancy Method*. Cambridge University Press, Cambridge 2000.

[6] Cohen, A., DeVore, R. A., Hochmuth, R., Restricted Nonlinear Approximation. *Constr. Approx.* **16** (2000), 85–113.

[7] Cordoba, A., Fernandez, P., Convergence and divergence of decreasing rearranged Fourier series. *SIAM J. Math. Anal.* **29** (1998), 1129–1139.

[8] DeVore, R. A., Nonlinear approximation. *Acta Numerica* (1998), 51–150.

[9] DeVore, R. A., Jawerth, B., Popov, V., Compression of wavelet decompositions. *Amer. J. Math.* **114** (1992), 737–785.

[10] DeVore, R. A., Konyagin, S. V., Temlyakov, V. N., Hyperbolic wavelet approximation. *Constr. Approx.* **14** (1998), 1–26.

[11] DeVore, R. A., Temlyakov, V. N., Nonlinear approximation by trigonometric sums. *J. Fourier Anal. Appl.* **2** (1995), 29–48.

[12] DeVore, R. A., Temlyakov, V. N., Some remarks on Greedy Algorithms. *Adv. Comput. Math.* **5** (1996), 173–187.

[13] Donahue, M., Gurvits, L., Darken, C., Sontag, E., Rate of convex approximation in non-Hilbert spaces. *Constr. Approx.* **13** (1997), 187–220.

[14] Dilworth, S. J., Kalton, N. J., Kutzarova, D., Temlyakov, V. N., The Tresholding Greedy Algorithm, Greedy Bases, and Duality. *Constr. Approx.* **19** (2003), 575–597.

[15] Dilworth, S. J., Kutzarova, D., Temlyakov, V. N., Convergence of some Greedy Algorithms in Banach spaces. *J. Fourier Anal. Appl.* **8** (2002), 489–505.

[16] Frazier, M., Jawerth, B., A discrete transform and decomposition of distribution spaces. *J. Funct. Anal.* **93** (1990), 34–170.

[17] Gluskin, E. D., Extremal properties of orthogonal parallelpipeds and their application to the geometry of Banach spaces. *Math. USSR-Sb.* **64** (1989), 85–96.

[18] Gribonval, R., Nielsen, M., Some remarks on non-linear approximation with Schauder bases. *East J. Approx.* **7** (2001), 267–285.

[19] Heinrich, S., Novak, E., Wasilkowski, G., Wozniakowski, H., The inverse of the star-discrepancy depends linearly on the dimension. *Acta Arith.* **96** (2001), 279–302.

[20] Ismagilov, R. S., Widths of sets in normed linear spaces and the approximation of functions by trigonometric polynomials. *Uspekhi Mat. Nauk* **29** (1974), 161–178.

[21] Kamont, A., Temlyakov, V. N., Greedy Approximation and the Multivariate Haar system. *Studia Math.* **161** (3) (2004), 199–223.

[22] Kashin, B. S., Widths of certain finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat.* **41** (1977), 334–351.

[23] Kashin, B. S., On approximation properties of complete orthonormal systems. *Trudy Mat. Inst. Steklov* **172** (1985), 187–191.

[24] Kashin, B. S., Temlyakov, V. N., On best m-term approximations and the entropy of sets in the space L^1. *Math. Notes* **56** (1994), 57–86.

[25] Kerkyacharian, G., Picard, D., Entropy, universal coding, approximation and bases properties. *Constr. Approx.* **20** (2004), 1–37.

[26] Kerkyacharian, G., Picard, D., Temlyakov, V.N., Some inequalities for the tensor product of greedy bases and weight-greedy bases. Manuscript, 2005, 1–14.

[27] Konyagin, S. V., Temlyakov, V. N., A remark on greedy approximation in Banach spaces. *East J. Approx.* **5** (1999), 1–15.

[28] Konyagin, S. V., Temlyakov, V. N., Greedy approximation with regard to bases and general minimal systems. *Serdica Math. J.* **28** (2002), 305–328.

[29] Konyagin, S. V., Temlyakov, V. N., Convergence of greedy approximation II. The trigonometric system. *Studia Math.* **159(2)** (2003), 161–184.

[30] Konyagin, S. V., Temlyakov, V. N., Convergence of Greedy Approximation for the trigonometric system. *Anal. Math.* **31** (2005), 85–115

[31] Körner, T. W., Divergence of decreasing rearranged Fourier series. *Ann. of Math.* **144** (1996), 167–180.

[32] Körner, T. W., Decreasing rearranged Fourier series. *J. Fourier Anal. Appl.* **5** (1999), 1–19.

[33] Kuipers, L., Niederreiter, H., *Uniform distribution of sequences*. Pure and Applied Mathematics, Wiley, New York, London, Sydney 1974.

[34] Lindenstrauss, J., Tzafriri, L., *Classical Banach Spaces I*. Ergebnisse Math Grenzgeb. 92, Springer-Verlag, Berlin 1977.

[35] Maiorov, V. E., Trigonometric diameters of the Sobolev classes W_p^r in the space L_q. *Math. Notes* **40** (1986), 590–597.

[36] Matoušek, J., *Geometric Discrepancy*. Algorithms Combin. 18, Springer-Verlag, Berlin 1999.

[37] Novak, E., Wozniakowski, H., When are Integration and Discrepancy tractable? In *Foundations of computational mathematics* (Oxford, 1999), London Math. Soc. Lecture Notes Ser. 284, Cambridge University Press, Cambridge 2001, 211–266.

[38] Temlyakov, V. N., Widths of Some Classes of Functions of Several Variables. *Soviet Math. Dokl.* **26** (1982), 619-622.

[39] Temlyakov, V. N., *Approximation of periodic functions*. Comput. Math. Anal. Ser., Nova Science Publishers, Inc., Commack, NY, 1993.

[40] Temlyakov, V. N., The best m-term approximation and Greedy Algorithms. *Adv. Comput. Math.* **8** (1998), 249–265.

[41] Temlyakov, V. N., Nonlinear m-term approximation with regard to the multivariate Haar system. *East J. Approx.* **4** (1998), 87–106.

[42] Temlyakov, V. N., Greedy Algorithm and m-term Trigonometric Approximation. *Constr. Approx.* **14** (1998), 569–587.

[43] Temlyakov, V. N., Greedy algorithms with regard to multivariate systems with special structure. *Constr. Approx.* **16** (2000), 399–425.

[44] Temlyakov, V. N., Greedy algorithms in Banach spaces. *Adv. Comput. Math.* **14** (2001), 277–292.

[45] Temlyakov, V. N., Universal bases and greedy algorithms for anisotropic function classes. *Constr. Approx.* **18** (2002), 529–550.

[46] Temlyakov, V. N., Nonlinear approximation with regard to bases. In *Approximation theory X* (St. Louis, MO, 2001), Innov. Appl. Math., Vanderbilt University Press, Nashville, TN, 2002, 373–402.

[47] Temlyakov, V. N., Nonlinear Methods of Approximation. *Found. Comput. Math.* **3** (2003), 33–107.

[48] Temlyakov, V. N., Cubature formulas and related questions. *J. Complexity* **19** (2003), 352–391.

[49] Temlyakov, V. N., Greedy Type Algorithms in Banach Spaces and Applications. *Constr. Approx.* **21** (2005), 257–292.

[50] Temlyakov, V. N., Weak Greedy Algorithms. *Adv. Comput. Math.* **12** (2000), 213–227.

[51] Tikhomirov, V. M., Widths of sets in function spaces and the theory of best approximations. *Uspekhi Mat. Nauk* **15** (1960), 81–120.

[52] Wojtaszczyk, P., On unconditional polynomial bases in L_p and Bergman spaces. *Constr. Approx.* **13** (1997), 1–15.

[53] Wojtaszczyk, P., Greedy algorithms for general systems, *J. Approx. Theory* **107** (2000), 293–314.

Department of Mathematics, University of South Carolina, Columbia, SC 29208, U.S.A.
E-mail: temlyak@math.sc.edu

Analytic capacity, rectifiability, and the Cauchy integral

Xavier Tolsa*

Abstract. A compact set $E \subset \mathbb{C}$ is said to be removable for bounded analytic functions if for any open set Ω containing E, every bounded function analytic on $\Omega \setminus E$ has an analytic extension to Ω. Analytic capacity is a notion that, in a sense, measures the size of a set as a non removable singularity. In particular, a compact set is removable if and only if its analytic capacity vanishes. The so-called Painlevé problem consists in characterizing removable sets in geometric terms. Recently many results in connection with this very old and challenging problem have been obtained. Moreover, it has also been proved that analytic capacity is semiadditive. We review these results and other related questions dealing with rectifiability, the Cauchy transform, and the Riesz transforms.

Mathematics Subject Classification (2000). Primary 30C85; Secondary 42B20, 28A75.

Keywords. Analytic capacity, rectifiability, Cauchy transform, Riesz transform, singular integrals.

1. Introduction

In this paper we survey recent results in connection with analytic capacity, rectifiability and the Cauchy and Riesz transforms. We are specially interested in the interaction between analytic and geometric notions. Most of the results that we will review are a mixture of harmonic analysis and geometric measure theory. Some of them may have also some little amount of complex analysis.

Let us introduce some notation and definitions. A compact set $E \subset \mathbb{C}$ is said to be removable for bounded analytic functions if for any open set Ω containing E, every bounded function analytic on $\Omega \setminus E$ has an analytic extension to Ω. In order to study removability, in the 1940s Ahlfors [Ah] introduced the notion of analytic capacity. The *analytic capacity* of a compact set $E \subset \mathbb{C}$ is

$$\gamma(E) = \sup |f'(\infty)|, \tag{1}$$

where the supremum is taken over all analytic functions $f : \mathbb{C} \setminus E \to \mathbb{C}$ with $|f| \leq 1$ on $\mathbb{C} \setminus E$, and $f'(\infty) = \lim_{z\to\infty} z(f(z) - f(\infty))$.

In [Ah], Ahlfors showed that E is removable for bounded analytic functions if and only if $\gamma(E) = 0$.

*Partially supported by grants MTM2004-00519 and Acción Integrada HF2004-0208 (Spain), and 2005-SGR-00744 (Generalitat de Catalunya).

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

Painlevé's problem consists in characterizing removable singularities for bounded analytic functions in a metric/geometric way. By Ahlfors' result this turns out to be equivalent to describing compact sets with positive analytic capacity in metric/geometric terms.

Vitushkin in the 1950s and 1960s showed that analytic capacity plays a central role in problems of uniform rational approximation on compact sets of the complex plane. Because of its applications to this type of problems he raised the question of the semiadditivity of γ. Namely, does there exist an absolute constant C such that

$$\gamma(E \cup F) \leq C(\gamma(E) + \gamma(F))\,?$$

It has recently been proved [To5] that analytic capacity is indeed semiadditive. Moreover, a characterization of removable sets for bounded analytic functions in terms of the so-called curvature of measures is also given in [To5]. In Section 2 of the present paper we will review these results and other recent advances in connection with analytic capacity and Painlevé's problem. We will describe some of the ideas involved in their proofs. In particular, we will see that L^2 estimates for the Cauchy transform play a prominent role in most of these results.

Recall that if ν is a finite complex Borel measure on $\mathbb{C}$, the *Cauchy transform* (or *Cauchy integral*) of ν is defined by

$$\mathcal{C}\nu(z) = \int \frac{1}{\xi - z} d\nu(\xi).$$

Although the integral above is absolutely convergent a.e. with respect to Lebesgue measure, it does not make sense, in general, for $z \in \operatorname{supp}(\nu)$. This is the reason why one considers the *truncated Cauchy transform* of ν, which is defined as

$$\mathcal{C}_\varepsilon \nu(z) = \int_{|\xi - z| > \varepsilon} \frac{1}{\xi - z} d\nu(\xi),$$

for any $\varepsilon > 0$ and $z \in \mathbb{C}$.

In Section 3 we will survey several results about the μ-almost everywhere (a.e.) existence of the principal value

$$\text{p.v.}\mathcal{C}\mu(z) = \lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon \mu(z),$$

where μ is some positive finite Borel measure on $\mathbb{C}$, and its relationship with rectifiability.

Section 4 deals with the natural generalization of analytic capacity to higher dimensions, the so-called Lipschitz harmonic capacity. The role played by the Cauchy transform in connection with analytic capacity corresponds to the Riesz transforms in the case of Lipschitz harmonic capacity. See Section 4 for more details.

In the final section of the paper we recall some open problems related to analytic capacity, the Cauchy and Riesz transforms, and rectifiability. This is a rather short list

which reflects our personal interests and it is not intended to be a complete account of open problems in the area.

Some comments about the notation used in the paper: as usual, the letter 'C' stands for an absolute constant which may change its value at different occurrences. The notation $A \lesssim B$ means that there is a positive absolute constant C such that $A \leq CB$. Also, $A \approx B$ is equivalent to $A \lesssim B \lesssim A$.

2. Analytic capacity and the Cauchy transform

2.1. Basic properties of analytic capacity. In a sense, analytic capacity measures the size of a set as a non removable singularity for bounded analytic functions. A direct consequence of the definition is that for all $\lambda \in \mathbb{C}$ and $E \subset \mathbb{C}$ compact one has $\gamma(\lambda + E) = \gamma(E)$ and $\gamma(\lambda E) = |\lambda|\gamma(E)$. Further, if E is connected, then

$$\operatorname{diam}(E)/4 \leq \gamma(E) \leq \operatorname{diam}(E).$$

The second inequality (which holds for any compact set E) follows from the fact that the analytic capacity of a ball coincides with its radius, and the first one is a consequence of Koebe's $1/4$ theorem (see [Gam, Chapter VIII] or [Gar2, Chapter I] for the details, for example). Thus if E is connected and different from a point, then it is non removable. This implies that any removable compact set must be totally disconnected.

The relationship between analytic capacity and Hausdorff measure is as follows:

- If $\dim_H(E) > 1$ (here $\dim_H$ stands for the Hausdorff dimension) then $\gamma(E) > 0$. This result is an easy consequence of Frostman's Lemma.
- $\gamma(E) \leq \mathcal{H}^1(E)$, where $\mathcal{H}^s$ is the s-dimensional Hausdorff measure, or length when $s = 1$. This follows from Cauchy's integral formula, and it was proved by Painlevé about one hundred years ago. In particular, notice that if $\dim_H(E) < 1$ then $\gamma(E) = 0$.

By the statements above, one infers that dimension 1 is the critical dimension in connection with analytic capacity. It turns out that some sets of positive length and dimension 1 have positive analytic capacity (for example, a segment), while others have vanishing analytic capacity. The latter assertion was proved by Vitushkin [Vi1]. Later on, an easier example of a set with positive length and zero analytic capacity was found by Garnett [Gar1] and Ivanov [Iv].

2.2. The Cauchy transform and the capacity γ_+. Recall that given a positive Borel measure μ on the complex plane, $\mathcal{C}\mu$ stands for the Cauchy transform of μ. If f is a μ-measurable function f on $\mathbb{C}$, we denote $\mathcal{C}_\mu f(z) := \mathcal{C}(f\,d\mu)(z)$ for $z \notin \operatorname{supp}(f)$, and $\mathcal{C}_{\mu,\varepsilon} f(z) := \mathcal{C}_\varepsilon(f\,d\mu)(z)$ for any $\varepsilon > 0$ and $z \in \mathbb{C}$. We say that $\mathcal{C}_\mu$ is bounded on $L^2(\mu)$ if the operators $\mathcal{C}_{\mu,\varepsilon}$ are bounded on $L^2(\mu)$ uniformly on $\varepsilon > 0$.

Let us denote by $M_+(\mathbb{C})$ the set of finite (positive) Borel measures on $\mathbb{C}$. The capacity γ_+ of a compact set $E \subset \mathbb{C}$ is

$$\gamma_+(E) := \sup\{\mu(E) : \mu \in M_+(\mathbb{C}), \operatorname{supp}(\mu) \subset E, \ \|\mathcal{C}\mu\|_{L^\infty(\mathbb{C})} \leq 1\}. \qquad (2)$$

That is, γ_+ is defined as γ in (1) with the additional constraint that f should coincide with $\mathcal{C}\mu$, where μ is some positive Borel measure supported on E (observe that $(\mathcal{C}\mu)'(\infty) = -\mu(\mathbb{C})$ for any Borel measure μ). Moreover, there is another slight difference: in (1) we required $\|f\|_{L^\infty(\mathbb{C}\setminus E)} \leq 1$, while in (2), $\|f\|_{L^\infty(\mathbb{C})} \leq 1$ (for $f = \mathcal{C}\mu$). Trivially, we have $\gamma_+(E) \leq \gamma(E)$.

We introduce now some additional notation. A Borel measure μ on $\mathbb{R}^d$ has growth of degree n if there exists some constant C such that $\mu(B(x, r)) \leq Cr^n$ for all $x \in \mathbb{R}^d$, $r > 0$. When $n = 1$, we say that μ has linear growth. If μ satisfies $\mu(B(x, r)) \approx r^n$ for all $x \in \operatorname{supp}(\mu)$, $0 < r \leq \operatorname{diam}(\operatorname{supp}(\mu))$, we say that μ is n-Ahlfors–David (n-AD) regular, or abusing language, AD regular. A set $E \subset \mathbb{C}$ is called n-AD regular (abusing language, AD regular) if $\mathcal{H}^n|_E$ is AD regular. We say that μ is doubling if there exists some constant C such that $\mu(B(z, 2r)) \leq C\mu(B(z, r))$ for all $z \in \operatorname{supp}(\mu)$, $r > 0$. In particular, AD regular measures are doubling.

The next theorem shows why L^2 estimates for the Cauchy transform are useful in connection with analytic capacity.

Theorem 2.1. *Let μ be a measure with linear growth on $\mathbb{C}$. Suppose that the Cauchy transform is bounded in $L^2(\mu)$. Then, for any compact $E \subset \mathbb{C}$ there exists a function h supported on E with $0 \leq h \leq 1$ such that $\int h\, d\mu \approx \mu(E)$, with $\|\mathcal{C}_\varepsilon(h\, d\mu)\|_\infty \leq C$ for all $\varepsilon > 0$, and $\|\mathcal{C}(h\, d\mu)\|_{\infty,\mathbb{C}\setminus E} \leq C$. All the constants depend only on the linear growth of μ and on $\|\mathcal{C}\|_{L^2(\mu),L^2(\mu)}$.*

In the statement above, $\|\mathcal{C}_\mu\|_{L^2(\mu),L^2(\mu)}$ stands for the operator norm of $\mathcal{C}_\mu$ on $L^2(\mu)$. That is, $\|\mathcal{C}_\mu\|_{L^2(\mu),L^2(\mu)} = \sup_{\varepsilon>0} \|\mathcal{C}_{\mu,\varepsilon}\|_{L^2(\mu),L^2(\mu)}$.

From the preceding result one infers that if E supports a non zero measure μ with linear growth such that the Cauchy transform $\mathcal{C}_\mu$ is bounded on $L^2(\mu)$, then $\gamma(E) \geq \gamma_+(E) > 0$.

Theorem 2.1 is from Davie and Øksendal [DØ] and it follows from the fact that the L^2 boundedness of the Cauchy transform implies weak (1, 1) estimates. Theorem 2.1 is obtained by a suitable dualization of these weak (1, 1) estimates. See [Uy] for a connected result prior to [DØ] which also involves a dualization of a weak (1, 1) inequality.

2.3. Menger curvature and rectifiability. Given three pairwise different points $x, y, z \in \mathbb{C}$, their *Menger curvature* is

$$c(x, y, z) = \frac{1}{R(x, y, z)},$$

where $R(x, y, z)$ is the radius of the circumference passing through x, y, z. If two among these points coincide, we let $c(x, y, z) = 0$. For a positive Borel measure μ,

we define the *curvature of* μ as

$$c^2(\mu) = \iiint c(x, y, z)^2 \, d\mu(x) d\mu(y) d\mu(z). \tag{3}$$

Given $\varepsilon > 0$, $c_\varepsilon^2(\mu)$ stands for the ε-truncated version of $c^2(\mu)$, defined as in the right hand side of (3), but with the triple integral over $\{(x, y, z) \in \mathbb{C}^3 : |x - y|, |y - z|, |x - z| > \varepsilon\}$.

The notion of curvature of a measure was introduced by Melnikov [Me] when he was studying a discrete version of analytic capacity, and it is one of the notions which is responsible of the big recent advances in connection with analytic capacity. The notion of curvature is connected to the Cauchy transform by the following result, proved by Melnikov and Verdera [MeV].

Proposition 2.2. *Let μ be a Borel measure on $\mathbb{C}$ with linear growth. We have*

$$\|\mathcal{C}_\varepsilon \mu\|_{L^2(\mu)}^2 = \frac{1}{6} c_\varepsilon^2(\mu) + O(\mu(\mathbb{C})), \tag{4}$$

where $|O(\mu(\mathbb{C}))| \leq C\mu(\mathbb{C})$.

Sketch of the proof of Proposition 2.2. If we do not worry about truncations and the absolute convergence of the integrals, we can write

$$\begin{aligned} \|\mathcal{C}\mu\|_{L^2(\mu)}^2 &= \int \left| \int \frac{1}{y - x} \, d\mu(y) \right|^2 d\mu(x) \\ &= \iiint \frac{1}{(y - x)(\overline{z - x})} \, d\mu(y) d\mu(z) d\mu(x). \end{aligned}$$

By Fubini (assuming that it can be applied correctly), permuting x, y, z, we get,

$$\|\mathcal{C}\mu\|_{L^2(\mu)}^2 = \frac{1}{6} \iiint \sum_{s \in S_3} \frac{1}{(z_{s_2} - z_{s_1})(\overline{z_{s_3} - z_{s_1}})} \, d\mu(z_1) d\mu(z_2) d\mu(z_3),$$

where S_3 is the group of permutations of three elements. An elementary calculation shows that

$$\sum_{s \in S_3} \frac{1}{(z_{s_2} - z_{s_1})(\overline{z_{s_3} - z_{s_1}})} = c(z_1, z_2, z_3)^2.$$

So we get

$$\|\mathcal{C}\mu\|_{L^2(\mu)}^2 = \frac{1}{6} \, c^2(\mu).$$

To argue rigorously, above we should use the truncated Cauchy transform $\mathcal{C}_\varepsilon \mu$ instead of $\mathcal{C}\mu$, and then we would obtain

$$\begin{aligned} \|\mathcal{C}_\varepsilon \mu\|_{L^2(\mu)}^2 &= \iiint_{\substack{|x-y|>\varepsilon \\ |x-z|>\varepsilon}} \frac{1}{(y - x)(\overline{z - x})} \, d\mu(y) d\mu(z) d\mu(x) \\ &= \iiint_{\substack{|x-y|>\varepsilon \\ |x-z|>\varepsilon \\ |y-z|>\varepsilon}} \frac{1}{(y - x)(\overline{z - x})} \, d\mu(y) d\mu(z) d\mu(x) + O(\mu(\mathbb{C})). \end{aligned} \tag{5}$$

By the linear growth of μ, it is easy to check that $|O(\mu(\mathbb{C}))| \leq \mu(\mathbb{C})$. As above, using Fubini and permuting x, y, z, one shows that the triple integral in (5) equals $c_\varepsilon^2(\mu)/6$. □

The identity (4) is remarkable because it relates an analytic notion (the Cauchy transform of a measure) with a metric/geometric one (curvature).

The above proposition was used in [MeV] to give a simple geometric proof of the L^2 boundedness of the Cauchy transform on a Lipschitz graph (the original proof is from Coifman, McIntosh and Meyer [CMM]). Indeed, using a Fourier type estimate, it was proved in [MeV] that if Γ is a Lipschitz graph, then $c^2(\mathcal{H}^1|_{\Gamma\cap B(z,r)} \lesssim r$ for all $z \in \Gamma$, $r > 0$. From (4) one infers that $\|\mathcal{C}_\varepsilon(b\,\chi_{B(z,r)}\mathcal{H}^1|_\Gamma)\|_{L^2(\mathcal{H}^1|\Gamma\cap B(z,r))} \lesssim \|b\|_\infty r^{1/2}$ for all $b \in L^\infty$, $z \in \Gamma$, and $r > 0$, uniformly on $\varepsilon > 0$. Now a simple argument shows that $\mathcal{C}_{\mathcal{H}^1|_\Gamma}$ sends boundedly L^∞ into $\mathrm{BMO}(\mathcal{H}^1|_\Gamma)$ and thus $H^1(\mathcal{H}^1|_\Gamma)$ into $L^1(\mathcal{H}^1|_\Gamma)$. Interpolating one gets the conclusion.

Let us turn our attention to rectifiability and its relationship with curvature of measures. A set is called rectifiable if it is $\mathcal{H}^1$-almost all contained in a countable union of rectifiable curves. On the other hand, it is called purely unrectifiable if it intersects any rectifiable curve at most in a set of zero length.

Now we wish to recall the traveling salesman theorem of P. Jones [Jo]. First we introduce some notation. Given $E \subset \mathbb{C}$ and a square Q, let V_Q be an infinite strip (or line in the degenerate case) of smallest possible width which contains $E \cap 3Q$, and let $w(V_Q)$ denote the width of V_Q. Then we set

$$\beta_E(Q) = \frac{w(V_Q)}{\ell(Q)},$$

where $\ell(Q)$ stands for the side length of Q. We denote by $\mathcal{D}$ the family of all dyadic squares in $\mathbb{C}$. In [Jo] the following result was proved:

Theorem 2.3. *A set $E \subset \mathbb{C}$ is contained in a rectifiable curve Γ (with finite length) if and only*

$$\sum_{Q\in\mathcal{D}} \beta_E(Q)^2 \ell(Q) < \infty. \tag{6}$$

Moreover, the length of the shortest curve Γ containing E satisfies

$$\mathcal{H}^1(\Gamma) \approx \mathrm{diam}(E) + \sum_{Q\in\mathcal{D}} \beta_E(Q)^2 \ell(Q),$$

with absolute constants.

The theorem also holds for sets E contained in $\mathbb{R}^d$. The proof of the "if" part of the theorem in [Jo] is also valid in this case. The "only if" part (for $\mathbb{R}^d$) was proved by Okikiolu [Ok]. Several versions of Jones' result which involve n-dimensional AD regular sets in $\mathbb{R}^d$ have been obtained by David and Semmes [DS1], [DS2]. In

fact, David and Semmes have developed a whole theory of the so-called "uniform rectifiability" for n-dimensional AD regular sets in $\mathbb{R}^d$.

The next result, proved by P. Jones (see [Pa, Chapter 3]), shows that there is a strong connection between curvature and the coefficients β in Jones' traveling salesman theorem.

Theorem 2.4. (a) *If $E \subset \mathbb{C}$ is 1-AD regular, then*

$$\sum_{Q \in \mathcal{D}} \beta_E(Q)^2 \ell(Q) \leq C\, c^2(\mathcal{H}^1|_E),$$

where C depends only on the AD regularity constant of E.

(b) *If μ is a measure with linear growth supported on a rectifiable curve $\Gamma \subset \mathbb{C}$, then*

$$c^2(\mu) \leq C \sum_{Q \in \mathcal{D}} \beta_\Gamma(Q)^2 \mu(Q),$$

where C depends only on the linear growth constant of μ.

From (a) in the preceding theorem and Theorem 2.3 it turns out that if $E \subset \mathbb{C}$ is AD regular and $c^2(\mathcal{H}^1|_E) < \infty$, then E is rectifiable. If one does not assume E to be AD regular, David and Léger [Lé] showed that the result still holds:

Theorem 2.5. *Let $E \subset \mathbb{C}$ be compact with $\mathcal{H}^1(E) < \infty$. If $c^2(\mathcal{H}^1|_E) < \infty$, then E is rectifiable.*

The proof of this result in [Lé] uses geometric techniques, in the spirit of the ones used by P. Jones for Theorem 2.3 in [Jo] and by David and Semmes in [DS1]. Recently, in [To9] a very different proof of Theorem 2.5 has been obtained. The new arguments are based on some kind of isoperimetric inequality involving analytic capacity and on the characterization of rectifiability in terms of densities.

From Theorem 2.5 and Proposition 2.2 one infers that if $\mathcal{H}^1(E) < \infty$ and the Cauchy transform is bounded on $L^2(\mathcal{H}^1|_E)$, then E must be rectifiable. A more quantitative version of this result proved by Mattila, Melnikov and Verdera [MMV] asserts that if E is AD regular and the Cauchy transform is bounded on $L^2(\mathcal{H}^1|_E)$, then E is contained in an AD regular curve Γ.

Recently, some of the results above dealing with rectifiability, β's, and curvature have been extended in different directions. For example, Lerman [Lr] has obtained a result analogous to Theorem 2.3 which involves very general Borel measures μ on $\mathbb{R}^d$ (instead of $\mathcal{H}^1|_E$) and $L^2(\mu)$ versions of Jones' β's. Ferrari, Franchi and Pajot [FFP] have extended the "if" part of the Theorem 2.3 to the Heisenberg group. Schul [Sch] has proved a version of the same theorem which is valid for Hilbert spaces. On the other hand, Hahlomaa [Hah] has obtained a version of Léger's Theorem 2.5 suitable for metric spaces.

2.4. Vitushkin's conjecture. For $\theta \in [0, \pi)$, let p_θ denote the orthogonal projection onto the line through the origin and direction $(\cos\theta, \sin\theta)$. Given a Borel set $E \subset \mathbb{C}$, its Favard length is

$$\mathrm{Fav}(E) = \int_0^\pi \mathcal{H}^1(p_\theta(E))\, d\theta. \tag{7}$$

Vitushkin conjectured in the 1960s that $\gamma(E) > 0$ if and only if $\mathrm{Fav}(E) > 0$. In 1986 Mattila [Ma1] showed that this conjecture is false. Indeed, he proved that the property of having positive Favard length is not invariant under conformal mappings while removability for bounded analytic functions remains invariant. Mattila's result did not tell which implication in the above conjecture was false. Later on, Jones and Murai [JM] constructed a set with zero Favard length and positive analytic capacity. An easier example using curvature was obtained more recently by Joyce and Mörters [JyM].

Although Vitushkin's conjecture is not true in full generality, it turns out that it holds in the particular case of sets with finite length. This was proved by G. David [Da] in 1998. Recall that when has E with finite length, by Besicovitch theorem, $\mathrm{Fav}(E) = 0$ if and only if E is purely unrectifiable.

The precise statement of David's result is the following.

Theorem 2.6. *Let $E \subset \mathbb{C}$ be compact with $\mathcal{H}^1(E) < \infty$. Then, $\gamma(E) = 0$ if and only if E is purely unrectifiable.*

This result is the solution of Painlevé's problem for sets with finite length. To be precise, let us remark that the "if" part of the theorem is not due to David. In fact, it follows from Calderón's theorem on the L^2 boundedness of the Cauchy transform on Lipschitz graphs with small Lipschitz constant and from Theorem 2.1. The "only if" part of the theorem, which is more difficult, is the one proved by David. Let us also mention that Mattila, Melnikov and Verdera [MMV] had proved previously the same result under the assumption that E is a 1-dimensional AD regular set.

The scheme of the proof of the "only if" part of Theorem 2.6 is the following. Let $E \subset \mathbb{C}$ be compact with $\gamma(E) > 0$ and finite length. Then there exists a function f analytic on $\mathbb{C} \setminus E$ such that $|f(z)| \leq 1$ on $\mathbb{C} \setminus E$ and $f'(\infty) = \gamma(E)$ (this is the so-called Ahlfors function, which maximizes $\gamma(E)$). Since $\mathcal{H}^1(E) < \infty$ it is not difficult to see that there exists some complex, bounded function g supported on E such that $f(z) = \mathcal{C}(g\, d\mathcal{H}^1|_E)(z)$ for $z \notin E$. Then it easily follows that $\|\mathcal{C}_\varepsilon(g\, d\mathcal{H}^1|_E)\|_{L^\infty} \leq C$ uniformly on $\varepsilon > 0$. On the other hand, g also satisfies $|\int g\, d\mathcal{H}^1|_E| = |f'(\infty)| = \gamma(E) > 0$. By a suitable $T(b)$ type theorem (which involves some delicate stopping time arguments, and non doubling measures) proved in [Da], one infers that there exists a subset $F \subset E$ with $\mathcal{H}^1(F) > 0$ such the Cauchy transform is bounded on $L^2(\mathcal{H}^1|_F)$. From Proposition 2.2 it follows that $c^2(\mathcal{H}^1|_F) < \infty$, and then by Theorem 2.5 F is rectifiable. So E cannot be purely unrectifiable.

Let us remark that the $T(b)$ theorem in [Da] uses a preliminary result from [DM]. A similar theorem had been previously obtained by Christ [Ch] in the AD regular case.

The result analogous to Theorem 2.6 for sets with infinite length is false. For this type of sets there is no such a nice geometric solution of Painlevé's problem, and we have to content ourselves with a characterization such as the one in Corollary 2.10 below (at least, for the moment).

2.5. Characterization of γ_+ in terms of curvature of measures and L^2 estimates for the Cauchy transform. The following theorem characterizes γ_+ in terms of curvature of measures and in terms of the L^2 norm of the Cauchy transform.

Theorem 2.7. *Let $\Sigma(E)$ denote class of Borel measures supported on E such that $\mu(B(x,r)) \leq r$ for all $x \in \mathbb{C}$, $r > 0$. For any compact set $E \subset \mathbb{C}$ we have*

$$\begin{aligned}\gamma_+(E) &\approx \sup\{\mu(E) :\ \mu \in \Sigma(E),\ c^2(\mu) \leq \mu(E)\} \\ &\approx \sup\{\mu(E) :\ \mu \in \Sigma(E),\ \|\mathcal{C}_\mu\|_{L^2(\mu),L^2(\mu)} \leq 1\}.\end{aligned} \tag{8}$$

Sketch of the proof of Theorem 2.7. Call S_1 and S_2 the first and second suprema on the right side of (8), respectively.

To see that $S_1 \gtrsim \gamma_+(E)$ take μ supported on E such that $\|\mathcal{C}\mu\|_\infty \leq 1$ and $\mu(E) \geq \gamma_+(E)/2$. One easily gets that $\|\mathcal{C}_\varepsilon \mu\|_\infty \lesssim 1$ on $\operatorname{supp}(\mu)$ for every $\varepsilon > 0$ and $\mu(B(x,r)) \leq Cr$ for all $r > 0$. From Proposition 2.2, it follows then that $c^2(\mu) \leq C\mu(E)$.

Consider now the inequality $S_2 \gtrsim S_1$. Let μ be supported on E with linear growth such that $c^2(\mu) \leq \mu(E)$ and $S_1 \leq 2\mu(E)$. We set

$$A := \left\{x \in E :\ \iint c(x,y,z)^2\, d\mu(y) d\mu(z) \leq 2\right\}.$$

By Tchebychev's inequality $\mu(A) \geq \mu(E)/2$. Moreover, for any set $B \subset \mathbb{C}$,

$$c^2(\mu|_{B\cap A}) \leq \iiint_{x\in B\cap A} c(x,y,z)^2\, d\mu(x) d\mu(y) d\mu(z) \leq 2\mu(B).$$

In particular, this estimate holds when B is any square in $\mathbb{C}$. Then, by the so-called $T(1)$ theorem (see [To1] or [NTV1]), one infers that $\mathcal{C}_{\mu|_A}$ is bounded on $L^2(\mu|_A)$. Thus $S_2 \gtrsim \mu(A) \approx S_1$.

Finally, the inequality $\gamma_+(E) \gtrsim S_2$ follows from Theorem 2.1. □

For the complete arguments of the preceding proof, see [To1] or [To4]. Notice that since the term

$$\sup\{\mu(E) :\ \mu \in \Sigma(E),\ \|\mathcal{C}_\mu\|_{L^2(\mu),L^2(\mu)} \leq 1\}$$

is countably semiadditive, from Theorem 2.7 one infers that γ_+ is also countably semiadditive.

Corollary 2.8. *The capacity γ_+ is countably semiadditive. That is, if E_i, $i = 1, 2, \ldots$, is a countable (or finite) family of compact sets, we have*

$$\gamma_+\Big(\bigcup_{i=1}^{\infty} E_i\Big) \leq C \sum_{i=1}^{\infty} \gamma_+(E_i).$$

2.6. Comparability between γ and γ_+. In [To5] the following result has been proved.

Theorem 2.9. *There exists an absolute constant C such that for any compact set $E \subset \mathbb{C}$ we have*

$$\gamma(E) \leq C\gamma_+(E).$$

As a consequence, $\gamma(E) \approx \gamma_+(E)$.

The comparability between γ and γ_+ had been previously proved by P. Jones for compact connected sets by geometric arguments, very different from the ones in [To5] (see [Pa, Chapter 3]). Also, in [MTV] it had already been shown that $\gamma \approx \gamma_+$ holds for a big class of Cantor sets. The proof of Theorem 2.9 in [To5] is inspired in part by the ideas in [MTV].

An obvious corollary of Theorem 2.9 and the characterization of γ_+ in terms of curvature in Theorem 2.7 is the following.

Corollary 2.10. *Let $E \subset \mathbb{C}$ be compact. Then, $\gamma(E) > 0$ if and only if E supports a non zero Borel measure with linear growth and finite curvature.*

Since we know that γ_+ is countably semiadditive, the same happens with γ:

Corollary 2.11. *Analytic capacity is countably semiadditive. That is, for any countable (or finite) family of compact sets E_i, $i = 1, 2, \ldots$, we have*

$$\gamma\Big(\bigcup_{i=1}^{\infty} E_i\Big) \leq C \sum_{i=1}^{\infty} \gamma(E_i).$$

Some few words about the proof of Theorem 2.7 are in order: it is enough to show that there exists some measure μ supported on E with linear growth, satisfying $\mu(E) \approx \gamma(E)$, and such that the Cauchy transform $\mathcal{C}_\mu$ is bounded on $L^2(\mu)$ with absolute constants. To this end, an important tool used in [To5] is the $T(b)$ theorem of Nazarov, Treil and Volberg in [NTV2], which is valid for non doubling measures. To apply this $T(b)$ theorem, one has to construct a suitable measure μ and a function $g \in L^\infty(\mu)$ fulfilling some precise conditions, similarly to the proof of Vitushkin's conjecture by David.

However, the situation now is more delicate because a direct application of that $T(b)$ theorem does not suffice. Indeed, let f be the Ahlfors function of E, so that f is analytic and bounded in $\mathbb{C} \setminus E$, with $f'(\infty) = \gamma(E)$. By a standard approximation

argument, it is not difficult to see that one can assume that $\mathcal{H}^1(E) < \infty$. Thus there exists some function g such that $f(z) = \mathcal{C}(g\, d\mathcal{H}^1|_E)(z)$ for $z \notin E$. If we argue like in [Da] (see Subsection 2.4 above), we will deduce that there exists a subset $F \subset E$ such that the Cauchy transform is bounded in $L^2(\mathcal{H}^1|_F)$. However, the size of F and the L^2 norm of the Cauchy transform $\mathcal{C}_{\mathcal{H}^1|_F}$ will depend strongly on the ratio $\mathcal{H}^1(E)/\gamma(E)$, which blows up as $\mathcal{H}^1(E) \to \infty$ or $\gamma(E) \to 0$. This difficulty is overcome in [To5] by using some ideas from potential theory and a suitable "induction on scales" argument.

Corollary 2.10 yields a characterization of removable sets for bounded analytic functions in terms of curvature of measures. Although this result has a definite geometric flavor, it is not clear if this is a really good geometric characterization. Nevertheless, in [To7] it has been shown that the characterization is invariant under bilipschitz mappings, using a corona type decomposition for non doubling measures. Previously, Garnett and Verdera [GV] had proved an analogous result for some Cantor sets. The problem about the behavior of removability and analytic capacity under bilipschitz mappings was raised by Verdera. See [Ve2].

2.7. Other results. In [To6], some results analogous to Theorems 2.7 and 2.9 have been obtained for the continuous analytic capacity α. This capacity is defined like γ in (1), with the additional requirement that the functions f considered in the sup should extend continuously to the whole complex plane. The capacity α is important because its many applications in connection with problems of uniform rational approximation in the complex plane, as shown by Vitushkin [Vi3]. In [To6] it is proved that α is semiadditive. This result has some nice consequences. For example, it implies the so-called *inner boundary conjecture*.

The inner boundary of a compact set $E \subset \mathbb{C}$, denoted by $\partial_i E$, is the set of boundary points of E which do not belong to the boundary of any connected component of $\mathbb{C} \setminus E$. The inner boundary conjecture (or theorem) says that if $\alpha(\partial_i E) = 0$, then any function analytic in $\overset{\circ}{E}$ and continuous on E can be approximated uniformly by functions which are analytic in neighborhoods of E (i.e. different neighborhoods for different functions).

The techniques for the proof of Theorem 2.9 have also been used by Prat [Pr] and Mateu, Prat and Verdera [MPV] to study the capacities γ_s associated to the s-dimensional signed Riesz kernel $k_s(x) = x/|x|^{s+1}$, with s non integer. Given a compact $E \subset \mathbb{R}^n$, the precise definition of $\gamma_s(E)$ is

$$\gamma_s(E) = \sup |\langle \nu, 1\rangle|, \tag{9}$$

where the supremum is taken over all distributions ν supported on E such that $K_s * \nu$ is an L^∞ function with $\|K_s * \nu\|_\infty \leq 1$.

The results in [Pr] and [MPV] show that the behavior of γ_s with s non integer is very different from the one with s integer. In [Pr] it is shown that sets with finite

s-dimensional Hausdorff measure have vanishing capacity γ_s when $0 < s < 1$. Moreover, for these s's it is proved in [MPV] that γ_s is comparable to the capacity $C_{\frac{2}{3}(n-s),\frac{3}{2}}$ from nonlinear potential theory. Recall that for $1 < p < \infty$ and $0 < tp \leq d$ the capacity $C_{t,p}$ of a compact set $E \subset \mathbb{R}^d$ is defined by

$$C_{t,p}(E) = \inf\bigl\{\|\varphi\|_p^p : \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d),\ \varphi * \tfrac{1}{|x|^{d-t}} \geq 1 \text{ on } E\bigr\}.$$

It is not known if the comparability $\gamma_s \approx C_{\frac{2}{3}(n-s),\frac{3}{2}}$ holds for non integers $s > 1$. This case seems much more difficult to study, although in the AD regular situation some results have been obtained in [Pr].

Using the corona type decomposition for measures with finite curvature and linear growth obtained in [To7], it has been proved in [To8] that if μ is a measure without atoms such that the Cauchy transform is bounded on $L^2(\mu)$, then any Calderón–Zygmund operator associated to an odd kernel sufficiently smooth is also bounded in $L^2(\mu)$.

3. Principal values for the Cauchy integral and related results

There is a strong relationship between rectifiability and the behavior of the Cauchy transform. Indeed, in Section 2 we saw that the L^2 boundedness of the Cauchy transform with respect to the arc length measure implies rectifiability. In this section we will describe some related results which involve the existence of principal values instead of L^2 boundedness.

Recall that given a Borel measure μ on $\mathbb{C}$, the principal of the Cauchy transform $\mathcal{C}\mu$ at $z \in \mathbb{C}$ is

$$\text{p.v.}\mathcal{C}\mu(z) = \lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon \mu(z),$$

whenever the limit exists. The maximal Cauchy transform of μ is

$$\mathcal{C}_*\mu(z) = \sup_{\varepsilon>0}\left|\int_{|\xi - z|>\varepsilon} \frac{1}{\xi - z}\, d\mu(\xi)\right|.$$

Obviously, the existence of $\text{p.v.}\mathcal{C}\mu(z)$ implies that $\mathcal{C}_*\mu(z)$ is finite. The converse needs not to be true.

Recall also that the upper and lower linear densities of μ at z are defined, respectively, by

$$\Theta_\mu^*(z) = \limsup_{r\to 0} \frac{\mu(B(z,r))}{2r}, \quad \Theta_{*,\mu}(z) = \liminf_{r\to 0} \frac{\mu(B(z,r))}{2r}.$$

When both densities coincide one writes $\Theta_\mu(z) := \Theta_\mu^*(z) = \Theta_{*,\mu}(z)$ and calls it linear density. A Dirac delta on a point $z \in \mathbb{C}$ is denoted by δ_z.

Mattila and Melnikov proved in [MMe] that if $E \subset \mathbb{C}$ has finite length and is rectifiable, then $\text{p.v.}\mathcal{C}\mu(z)$ exists $\mathcal{H}^1$-a.e. on E, for any Borel measure μ (see [Ve1]

for an easier proof). Using this result and Léger's Theorem 2.5 it was shown later in [To3] that if σ is a measure with linear growth such that the Cauchy transform is bounded on $L^2(\sigma)$, then p.v.$\mathcal{C}\mu(z)$ exists σ-a.e. As a consequence, by Theorem 2.9, for any Borel measure μ on $\mathbb{C}$, p.v.$\mathcal{C}\mu(z)$ exists γ-a.e., that is to say, the set where p.v.$\mathcal{C}\mu(z)$ fails to exist has zero analytic capacity (this result answers a question from Verdera [Ve2]).

In the converse direction (existence of principal values implies rectifiability) the first result was obtained by Mattila [Ma3]: he proved that if $\Theta_{*,\mu}(z) > 0$ and p.v.$\mathcal{C}\mu(z)$ exists (and is finite) for μ-a.e. $z \in \mathbb{C}$, then μ is concentrated on a rectifiable set, that is μ vanishes out of a rectifiable set. Let us remark that this result was obtained before the proof of the identity (4) which relates the Cauchy transform and curvature. Mattila's techniques were based on the use of tangent measures. In [Huo], Huovinen extended Mattila's result to other kernels in the plane different from the Cauchy transform.

In [Ma3] Mattila wondered if the assumption $\Theta_{*,\mu}(z) > 0$ μ-a.e. might be replaced by $\Theta^*_\mu(z) > 0$ μ-a.e. In [To2] a partial answer was given: using the $T(b)$ theorem of Nazarov, Treil and Volberg in [NTV2] and the "curvature method" (identity (4) and Léger's Theorem 2.5 were used), it was shown that the main result in [Ma3] also holds for measures such that $0 < \Theta^*_\mu(z) < \infty$ μ-a.e. Also, it was proved that one can replace the assumption on the existence of principal values by finiteness of the maximal Cauchy transform. As a consequence, one deduces that if $E \subset \mathbb{C}$ has finite length and $\mathcal{C}_*\mathcal{H}^1|_E(z) < \infty$ $\mathcal{H}^1$-a.e. on E, then E must be rectifiable.

A complete answer to Mattila's question has been given in [To10] recently. The precise result is the following.

Theorem 3.1. *Assume that μ is a finite Radon measure on the complex plane and set $E = \{z \in \operatorname{supp}(\mu) : \mathcal{C}_*\mu(z) < \infty\}$. Then $\mu|_E$ can be decomposed as $\mu|_E = \mu_d + \mu_r + \mu_0$, where*

$$\mu_d = \sum_i a_i \delta_{z_i}$$

for some $a_i > 0$ and $z_i \in \mathbb{C}$,

$$\mu_r = \sum_i g_i \mathcal{H}^1|_{\Gamma_i}$$

for some rectifiable curves Γ_i and non negative functions $g_i \in L^1(\mathcal{H}^1|_{\Gamma_i})$, and

$$\mu_0 = \sum_i \sigma_i,$$

where σ_i are measures with finite curvature such that $\Theta_{\sigma_i}(z) = 0$ for σ_i-a.e. $z \in \mathbb{C}$.

Notice that the preceding result asserts that there is a kind of "dimensional gap" between 0 and 1 for the measures μ such that $\mathcal{C}_*\mu(z) < \infty$ μ-a.e. For instance, for $0 < s < 1$ there are no measures of the form $\mu = \mathcal{H}^s|_E$, with $0 < \mathcal{H}^s(E) < \infty$, such that $\mathcal{C}_*\mu(z) < \infty$ μ-a.e.

A straightforward corollary of the theorem above is the following.

Corollary 3.2. *Let μ be a finite Radon measure on the complex plane such that $\Theta^*_\mu(z) > 0$ for μ-a.e. $z \in \mathbb{C}$. If $\mathcal{C}_*\mu(z) < \infty$ at μ-a.e. $z \in \mathbb{C}$, then μ can be decomposed as $\mu = \mu_d + \mu_r$, where $\mu_d = \sum_i a_i \delta_{z_i}$ for some $a_i > 0$ and $z_i \in \mathbb{C}$, and $\mu_r = \sum_i g_i \mathcal{H}^1|_{\Gamma_i}$ for some rectifiable curves Γ_i and non negative functions $g_i \in L^1(\mathcal{H}^1|_{\Gamma_i})$.*

In particular, under the assumptions of the corollary, μ is concentrated on a countable union of rectifiable curves.

The main difficulty to prove Theorem 3.1 consists in proving that if $\mathcal{C}_*\mu(z) < \infty$ μ-a.e. on $\mathbb{C}$, then on the set $\{z : \Theta^*_\mu(z) = \infty\}$ μ must be discrete (i.e. the addition of countably many point masses). Once this is proved, one can argue as in [To2]. A basic tool for the proof is the identity (4) again. However, notice that (4) holds for measures with linear growth, and the measures μ considered in Theorem 3.1 may be very far from having this property. This is the main obstacle that is overcome in [To10].

Before [To10], Jones and Poltoratski proved in [JP], among other things, that if μ is supported on a line (and more generally on a $\mathcal{C}^1$ curve) and $\mathcal{C}_*\mu(z) < \infty$ μ-a.e., then μ equals a countable collection of point masses plus some measure absolutely continuous with respect to arc length. Observe that this result is implied by Corollary 3.2, because any measure μ supported on a line satisfies $\Theta^*_\mu(z) > 0$ μ-a.e. In [JP] it was also shown that if one does not assume μ to be supported on a line and instead one asks the same conditions as Mattila in [Ma3] (i.e. $\Theta_{*,\mu}(z) > 0$ and p.v.$\mathcal{C}\mu(z)$ exists μ a.e. on $\mathbb{C}$), then the conclusion of Corollary 3.2 holds: μ equals a countable collection of point masses plus some measure absolutely continuous with respect to arc length on a rectifiable set.

The same techniques used for Theorem 3.1 also yield the following result, proved previously by Jones and Poltoratski [JP] when μ is supported on a $\mathcal{C}^1$ curve.

Theorem 3.3. *Assume that μ is a finite Radon measure on the complex plane and set $E = \{x \in \operatorname{supp}(\mu) : \mathcal{C}_\varepsilon\mu(x) = o(\mu(B(x,\varepsilon)/\varepsilon)$ as $\varepsilon \to 0+\}$. Then $\mu|_E$ can be decomposed as $\mu|_E = \mu_d + \mu_r + \mu_0$, with μ_d, μ_r, and μ_0 as in Theorem* 3.1.

The arguments in [To10] rely on the relationship between the Cauchy transform and curvature and so they do not extend to higher dimensions (i.e. to Riesz transforms). This is not the case with the results in [Ma3] and [JP]. Mattila's results in [Ma3] have been extended to the case of Riesz transforms by Mattila and Preiss [MPr], while some theorems in [JP] deal both with Cauchy and Riesz transforms. See next section for the precise definition of Riesz transforms.

4. Lipschitz harmonic capacity and Riesz transforms

Let $E \subset \mathbb{R}^d$ be a compact set. Its Lipschitz harmonic capacity is

$$\kappa(E) = \sup |\langle \Delta f, 1\rangle|,$$

where Δ stands for the Laplacian in the distributional sense and the supremum is taken over all functions f which are harmonic in $\mathbb{R}^d \setminus E$ and Lipschitz on $\mathbb{R}^d$, with $\|\nabla f\|_\infty \leq 1$ and $\nabla f(\infty) = 0$. If, in addition, in the supremum above one asks f to be $\mathcal{C}^1$, then one obtains the so-called $\mathcal{C}^1$ harmonic capacity.

Given a distribution ν on $\mathbb{R}^d$, let $R(\nu) = \frac{y}{|y|^d} * \nu$ be its (vectorial) Riesz transform, so that when ν is a real measure,

$$R\nu(x) = \Big(\frac{y}{|y|^d} * \nu\Big)(x) = \int \frac{x-y}{|x-y|^d}\, d\nu(y). \tag{10}$$

If μ is a positive Borel measure and f a μ-measurable function, we consider the operator $R_\mu f := R(f\, d\mu)$. For $\varepsilon > 0$, the truncated Riesz transforms R_ε, $R_{\mu,\varepsilon}$ are defined analogously to $\mathcal{C}_\varepsilon$ and $\mathcal{C}_{\mu,\varepsilon}$. One says that R_μ is bounded on $L^2(\mu)$ if the operators $R_{\mu,\varepsilon}$ are bounded on $L^2(\mu)$ uniformly on $\varepsilon > 0$.

An equivalent definition of $\kappa(E)$ in terms of Riesz transforms is the following:

$$\kappa(E) = \sup \sigma_d^{-1} |\langle \nu, 1\rangle|, \tag{11}$$

where σ_d stands for the $(d-1)$-dimensional volume of the sphere $\{|x| = 1\}$ in $\mathbb{R}^d$ and the supremum is taken over all real distributions ν supported on E such that $R\nu$ is a bounded function on $\mathbb{R}^d$, with $\|R\nu\|_\infty \leq 1$. An analogous definition exists for the $\mathcal{C}^1$ harmonic capacity.

The notions of Lipschitz harmonic capacity and $\mathcal{C}^1$ harmonic capacity were introduced by Paramonov [Par] in order to study problems of approximation of harmonic functions in the $\mathcal{C}^1$ norm. These capacities can be considered as real versions of the analytic capacity γ and the continuous analytic capacity α in $\mathbb{R}^d$, $d \geq 2$, respectively. Indeed, in $\mathbb{R}^2$, κ coincides (modulo a multiplicative absolute constant) with the so-called real analytic capacity $\gamma_{\mathbb{R}}$:

$$\gamma_{\mathbb{R}}(E) = \sup |f'(\infty)|,$$

where the supremum is taken over all functions f analytic on $\mathbb{C} \setminus E$ with $\|f\|_\infty \leq 1$ which are Cauchy transforms of *real* distributions. Analogously with respect to the $\mathcal{C}^1$ harmonic capacity.

Some geometric properties of κ in $\mathbb{R}^d$ have been studied in [MPa]. In particular, the relationship with the $(d-1)$-dimensional Hausdorff measure. Analogously to the case of analytic capacity, $(d-1)$-dimensional rectifiability seems to play a key role in the understanding of κ. Recall that $E \subset \mathbb{R}^d$ is called n-rectifiable if there are

Lipschitz maps $f_i : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$\mathcal{H}^n\Big(E \setminus \bigcup_{i=1}^{\infty} f_i(\mathbb{R}^n)\Big) = 0.$$

For example, if $\mathcal{H}^{d-1}(E) > 0$ and E is $(d-1)$-rectifiable, then $\kappa(E) > 0$. However, unlike in the case of analytic capacity, it is not known if the compact sets $E \subset \mathbb{R}^d$ with $\mathcal{H}^{d-1}(E) < \infty$ and $\kappa(E) = 0$ must be purely $(d-1)$-unrectifiable (recall Theorem 2.6). The main obstacle is the lack of a formula similar to (4) which relates the L^2 norm of the Riesz transforms with something which has some geometric meaning, or at least with something non negative, which may act as a square function for Riesz transforms. See [Fa] for this question.

Because of the same reason, many of the results in connection with the Cauchy transform which have been described in Sections 2 and 3 are missing in the case of Riesz transforms. See next section for some open questions.

On the other hand, although in the proof of the comparability between γ and γ_+ in [To5], curvature seems to play an important role, Volberg [Vo] has been able to show κ and κ_+ are also comparable (κ_+ is defined like κ in (11), but taking only the supremum over positive measures ν supported on E). As a consequence, κ is countably semiadditive, because κ_+ has the following characterization (compare with Theorem 2.7):

$$\kappa_+(E) \approx \sup\{\mu(E) : \operatorname{supp}(\mu) \subset E,\ \mu(B(x,r)) \leq r^{d-1} \text{ for all } x, r, \\ \|R_\mu\|_{L^2(\mu),L^2(\mu)} \leq 1\},$$

which is clearly subadditive. The main difference between the arguments used in [To5] for analytic capacity and ones in [Vo] for Lipschitz harmonic capacity stems from the choice of an appropriate potential useful for κ_+ which is very different from the one used for γ_+ in [To5].

Another paper concerning Lipschitz harmonic capacity is [MT]. In this article some Cantor sets in $\mathbb{R}^d$ are considered, and their Lipschitz harmonic capacity is estimated. The results are analogous to the ones in [MTV] for analytic capacity. The main difficulty in [MT] consists in estimating the L^2 norm of the Riesz transforms with respect to the natural probability measure associated to these Cantor sets.

5. Some open problems

In this section we collect some open problems related to analytic capacity, the Cauchy and Riesz transforms, and rectifiability. Most of them are well known, and there is no attempt at originality or completeness.

1. Subadditivity of analytic capacity. Is analytic capacity subadditive? That is to say, we are asking if γ is semiadditive with constant 1:

$$\gamma(E \cup F) \leq \gamma(E) + \gamma(F) \quad \text{for all compact sets } E, F \subset \mathbb{C}.$$

There are a couple of facts that suggest that this may be true: first, γ is *countably* semiadditive; and second, the above inequality holds when E and F are disjoint compact connected sets [Su].

Even when $\gamma(F) = 0$, the identity $\gamma(E \cup F) = \gamma(E)$ remains unproved.

2. The Cauchy capacity. Given $E \subset \mathbb{C}$ compact, consider the so-called *Cauchy capacity* of E:

$$\gamma_C(E) = \sup |\nu(E)|,$$

where the supremum is taken over all *complex measures* supported on E such that $|\mathcal{C}\nu(z)| \leq 1$ for $z \notin E$. This definition is similar to the one of analytic capacity, but now the supremum is restricted to Cauchy transforms of complex measures instead of distributions. Is it true that

$$\gamma(E) = \gamma_C(E)?$$

Notice that Theorem 2.9 implies that $\gamma(E) \approx \gamma_C(E)$. However, it is not known if the identity holds, except in some particular cases: for example, when E has finite length.

There exist compact sets E and functions f that are bounded and analytic on $\mathbb{C} \setminus E$ which are not the Cauchy transform of any complex measure ν. Nevertheless, the identity $\gamma(E) = \gamma_C(E)$ may still hold. See [Kha1] and [Kha2] for more information about this question.

3. Analytic capacity and Favard length. Recall the definition of Favard length in (7). We have already mentioned in Subsection 2.4 that Vitushkin's conjecture fails for sets with infinite length [Ma1]. In fact, Jones and Murai [JM] showed that there are sets with zero Favard length and positive analytic capacity.

So one of the implications in Vitushkin's conjecture is false. However the other implication is still open:

$$\text{Fav}(E) > 0 \quad \Rightarrow \quad \gamma(E) > 0?$$

Observe that by the characterization of γ in terms of curvature, this question can be restated in a more geometric way.

A related problem is the following. Let E be the so-called corner quarters Cantor set. This set is constructed as follows. Let $Q^0 = [0, 1] \times [0, 1]$. At the first step we take four closed squares inside Q^0 with side length $1/4$, with sides parallel to the coordinate axes, and so that each square contains a vertex of Q^0. At the second step we apply the preceding procedure to each of the four squares obtained in the first step,

so that we get 16 squares of side length $1/16$. Proceeding inductively, at each step we obtain 4^n squares Q_j^n, $j = 1, \ldots, 4^n$ with side length 4^{-n}. We denote

$$E_n = \bigcup_{j=1}^{4^n} Q_j^n,$$

and we define $E = \bigcap_{n=1}^{\infty} E_n$. This set has positive finite length and is purely unrectifiable, and so it has zero analytic capacity by David's theorem (this had been proved before David's result in [Gar1] and [Iv]). In fact, by [MTV] one has the asymptotic estimate $\gamma(E_n) \approx n^{-1/2}$. By Besicovitch theorem, we have $\mathrm{Fav}(E) = 0$. However, the asymptotic behavior of $\mathrm{Fav}(E_n)$ as $n \to \infty$ is not known. An interesting problem consists in finding more or less precise estimates for the asymptotic behavior of $\mathrm{Fav}(E_n)$.

From some results due to Mattila [Ma2] one gets the lower estimate $\mathrm{Fav}(E_n) \gtrsim 1/n$. Other recent results by Peres and Solomyak [PS] for random Cantor sets suggest that the estimate $\mathrm{Fav}(E_n) \approx 1/n$ may hold. However, up to now the best upper estimate is

$$\mathrm{Fav}(E_n) \lesssim \exp(-a \log_* n) \tag{12}$$

where C, a are positive absolute constants and $\log_*$ is the function

$$\log_* x = \min\Big\{n \in \mathbb{N} : \underbrace{\log \log \cdots \log}_{n} x \leq 1\Big\}.$$

Inequality (12) has been obtained in [PS]. It turns out that $\exp(-a \log_* n) \to 0$ extremely slowly, much more slowly than $1/n$.

Many difficulties that arise in connection with Favard length are related to the fact that there is no a *quantitative* proof of Besicovitch theorem. For more information concerning Favard length and projections we recommend to have a look at the nice survey [Ma4].

4. Vanishing Cauchy transforms. Let ν be a complex Borel measure on $\mathbb{C}$. Suppose that $\mathrm{p.v.}\mathcal{C}\nu(z)$ exists and vanishes $|\nu|$-a.e. on $\mathbb{C}$. Does this imply that ν is an atomic measure?

Notice that if ν is a positive measure, then Theorem 3.1 applies to ν and so in this case we know that $\nu = \nu_d + \nu_r + \nu_0$, where ν_d is discrete and ν_r, ν_0 are as in Theorem 3.1.

In [TVe] two particular cases have been studied. In the first one ν is absolutely continuous with respect to Lebesgue measure, and in the second one ν is a positive measure with linear growth and finite curvature. In both cases, if $\mathrm{p.v.}\mathcal{C}\nu(z)$ exists and vanishes $|\nu|$-a.e., then $\nu = 0$.

Let us remark that there are positive discrete (non zero) measures such that $\mathrm{p.v.}\mathcal{C}\nu(z)$ exists and vanishes ν-a.e. A trivial example is a single point mass. One can also construct other examples with countably many point masses.

5. Riesz transforms and rectifiability. Let $E \subset \mathbb{R}^d$ be a compact set with $0 < \mathcal{H}^n(E) < \infty$, for some integer $0 < n < d$. Take $\mu = \mathcal{H}^n|_E$ and consider the n-dimensional Riesz transform:

$$R_\mu^n f(x) = \int \frac{x-y}{|x-y|^{n+1}} f(y)\, d\mu(y),$$

for $f \in L^2(\mu)$, and $x \notin E$. As usual, we say that R_μ^n is bounded in $L^2(\mu)$ if the corresponding ε-truncated operators are bounded in $L^2(\mu)$ uniformly on $\varepsilon > 0$. If R_μ^n is bounded in $L^2(\mu)$, is then E n-rectifiable? The answer is known (and it is positive in this case) only for $n = 1$, because the curvature method works for $n = 1$. By [Vo], when $n = d - 1$ this question is equivalent to the following, which has already appeared in Section 4: is it true that $\kappa(E) = 0$ if and only if E is purely $(d-1)$-unrectifiable?

A variant of this problem consist in taking E AD regular and n-dimensional. If R_μ^n is bounded in $L^2(\mu)$, is then E uniformly n-rectifiable? For the definition of uniform rectifiability, see [DS1] and [DS2] (for the reader's convenience let us say that, roughly speaking, uniform rectifiability is the same as rectifiability plus some quantitative estimates). For $n = 1$ the answer is true again, because of curvature. The result is from Mattila, Melnikov and Verdera [MMV]. For $n > 1$, in [DS1] and [DS2] some partial answers are given. Let H_n be class of all the operators T defined as follows:

$$Tf(x) = \int k(x-y) f(y)\, d\mu(x),$$

where k is some odd kernel (i.e. $k(-x) = -k(x)$) smooth away from the origin such that $|x|^{n+j}|\nabla^j k(x)| \in L^\infty(\mathbb{R}^d \setminus \{0\})$ for $j \geq 0$. Suppose that all operators T from H_n are bounded in $L^2(\mu)$. Then it is shown in [DS1] that E is uniformly rectifiable. See [DS2] for other related results.

Consider again the case of a general compact set $E \subset \mathbb{R}^d$ with $0 < \mathcal{H}^n(E) < \infty$, for some integer $0 < n < d$, and set $\mu = \mathcal{H}^n|_E$. If μ satisfies

$$\liminf_{r \to 0} \frac{\mu(B(x,r))}{r^n} > 0 \qquad \mu\text{-a.e. on } \mathbb{R}^d,$$

then the μ-a.e. existence of the principal value $\lim_{\varepsilon \to 0} R_{\mu,\varepsilon}^n 1(x)$ implies that E is n-rectifiable, by a theorem of Mattila and Preiss [MPr]. However, this does not help to solve the questions above because it is not known if the $L^2(\mu)$ boundedness of the Riesz transforms R_μ^n implies the existence of principal values. Notice the contrast with the case of the Cauchy transform (see Section 3), where the latter assertion is known to be true, because of curvature again.

I would like to thank Joan Verdera for his remarks on preliminary versions of this paper.

References

[Ah] Ahlfors, L. V., Bounded analytic functions. *Duke Math. J.* **14** (1947), 1–11.

[Ch] Christ, M., A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral. *Colloq. Math.* **60/61** (2) (1990), 601–628.

[CMM] Coifman, R. R., McIntosh, A., and Meyer, Y., L'integrale de Cauchy définit un opérateur borné sur L^2 pour les courbes lipschitziennes. *Ann. of Math.* (2) **116** (1982), 361–387.

[Da] David, G., Unrectifiable 1-sets have vanishing analytic capacity. *Rev. Mat. Iberoamericana* **14** (2) (1998), 369–479.

[DM] David, G., and Mattila, P., Removable sets for Lipschitz harmonic functions in the plane. *Rev. Mat. Iberoamericana* **16** (1) (2000), 137–215.

[DS1] David, G., and Semmes, S., Singular integrals and rectifiable sets in R_n: Au-delà des graphes lipschitziens. *Astérisque* **193** (1991).

[DS2] David, G., and Semmes, S., *Analysis of and on uniformly rectifiable sets*. Math. Surveys Monogr. 38, Amer. Math. Soc., Providence, RI, 1993.

[DØ] Davie, A. M., and Øksendal, B., Analytic capacity and differentiability properties of finely harmonic functions. *Acta Math.* **149** (1–2) (1982), 127–152.

[Fa] Farag, H., The Riesz kernels do not give rise to higher-dimensional analogues of the Menger-Melnikov curvature. *Publ. Mat.* **43** (1) (1999), 251–260.

[FFP] Ferrari, F., Franchi, B., and Pajot, H., The Geometric Traveling Salesman Problem in the Heisenberg Group. Preprint, 2005.

[Gam] Gamelin, T., *Uniform Algebras*. Prentice Hall, Englewood Cliffs, N.J., 1969.

[Gar1] Garnett, J. B., Positive length but zero analytic capacity. *Proc. Amer. Math. Soc.* **24** (1970), 696–699.

[Gar2] Garnett, J., *Analytic capacity and measure*. Lecture Notes in Math. 297, Springer-Verlag, Berlin 1972.

[Hah] I. Hahlomaa, I., Menger curvature and Lipschitz parametrizations in metric spaces. *Fund. Math.* **185** (2) (2005), 143–169.

[Huo] Huovinen, P., Singular integrals and rectifiability of measures in the plane. Dissertation, University of Jyväskylä, Jyväskylä, 1997. Ann. Acad. Sci. Fenn. Math. Diss. No. 109, 1997, 63 pp.

[Iv] Ivanov, L. D., On sets of analytic capacity zero. In *Linear and Complex Analysis Problem Book 3* (Part II), Lectures Notes in Math. 1574, Springer-Verlag, Berlin 1994, 150–153.

[GV] Garnett J., and Verdera, J., Analytic capacity, bilipschitz maps and Cantor sets. *Math. Res. Lett.* **10** (2003), 515–522.

[Jo] Jones, P. W., Rectifiable sets and the traveling salesman problem. *Invent. Math.* **102** (1990), 1–15.

[JM] Jones, P. W., and Murai, T., Positive analytic capacity but zero Buffon needle probability. *Pacific J. Math.* **133** (1988), 89–114.

[JP] Jones, P. W., and Poltoratski, A. G., Asymptotic growth of Cauchy transforms. *Ann. Acad. Sci. Fenn. Math.* **29** (1) (2004), 99–120.

[JyM] Joyce, H., and Mörters, P., A set with finite curvature and projections of zero length. *J. Math. Anal. Appl.* **247** (2000), 126–135.

[Kha1] Khavinson, Ya, S., Golubev sums: a theory of extremal problems that are of the analytic capacity problem type and of accompanying approximation processes. *Uspekhi Mat. Nauk* **54** (4) (1999), 75–142; English transl. *Russian Math. Surveys* **54** (4) (1999), 753–818.

[Kha2] Khavinson, Ya, S., Duality relations in the theory of analytic capacity. *Algebra i Analiz* 15 (2003), no. 1, 3–62; English transl. *St. Petersburg Math. J.* **15** (1) (2004), 1–40.

[Lé] Léger, J. C. , Menger curvature and rectifiability. *Ann. of Math.* **149** (1999), 831–869.

[Lr] Lerman, G., Quantifying curvelike structures of measures by using L^2 Jones quantities. *Comm. Pure Appl. Math.* **56** (2003), 1294–1365.

[MPV] Mateu, J., Prat, L., and Verdera, J., The capacities associated to signed Riesz kernels, and Wolff potentials. *J. Reine Angew. Math.* **578** (2005), 201–223.

[MT] Mateu, J., and Tolsa, X., Riesz transforms and harmonic Lip_1-capacity in Cantor sets. *Proc. London Math. Soc.* **89** (3) (2004), 676–696.

[MTV] Mateu, J., Tolsa, X., and Verdera, J., The planar Cantor sets of zero analytic capacity and the local $T(b)$-Theorem. *J. Amer. Math. Soc.* **16** (2003), 19–28.

[Ma1] Mattila, P., Smooth maps, null sets for integralgeometric measure and analytic capacity. *Ann. of Math.* **123** (1986), 303–309.

[Ma2] Mattila, P., Orthogonal projections, Riesz capacities, and Minkowski content. *Indiana Univ. Math. J.* **39** (1) (1990), 185–198.

[Ma3] Mattila, P., Cauchy Singular Integrals and Rectifiability of Measures in the Plane. *Adv. Math.* **115** (1995), 1–34.

[Ma4] Mattila, P., Hausdorff dimension, projections, and Fourier transform. *Publ. Mat.* **48** (2004), 3–48.

[MMe] Mattila, P., and Melnikov, M. S., Existence and weak type inequalities for Cauchy integrals of general measures on rectifiable curves and sets. *Proc. Amer. Math.* **120** (1994), 143–149.

[MMV] Mattila, P., Melnikov, M. S., and Verdera, J., The Cauchy integral, analytic capacity, and uniform rectifiability. *Ann. of Math.* (2) **144** (1996), 127–136.

[MPa] Mattila, P., and Paramonov, P. V., On geometric properties of harmonic Lip_1-capacity. *Pacific J. Math.* **171** (2) (1995), 469–490.

[MPr] Mattila, P., and Preiss, D., Rectifiable measures in $\mathbb{R}^n$ and existence of principal values for singular integrals. *J. London Math. Soc.* (2) **52** (3) (1995), 482–496.

[Me] Melnikov, M. S., Analytic capacity: discrete approach and curvature of a measure. *Mat. Sb.* **186** (6) (1995), 57–76; English transl. *Math. Sb.* **186** (6) (1995), 827–846.

[MeV] Melnikov, M. S., and Verdera, J., A geometric proof of the L^2 boundedness of the Cauchy integral on Lipschitz graphs. *Internat. Math. Res. Notices* **(1995)** (7) (1995), 325–331.

[NTV1] Nazarov, F., Treil, S., and Volberg, A., Cauchy integral and Calderón-Zygmund operators in non-homogeneous spaces. *Internat. Res. Math. Notices* **1997** (15) (1997), 703–726.

[NTV2] Nazarov, F., Treil, S., and Volberg, A., The $T(b)$ theorem on non-homogeneous spaces that proves a conjecture of Vitushkin. Preprint n. 519, Centre de Recerca Matemàtica, Barcelona 2002.

[Ok] Okikiolu, K., Characterization of subsets of rectifiable curves in R^n. *J. London Math. Soc.* (2) **46** (1992), 336–348.

[Pa] Pajot, H., *Analytic capacity, rectifiability, Menger curvature and the Cauchy integral*. Lecture Notes in Math. 1799, Springer-Verlag, Berlin 2002.

[Par] Paramonov, P. V., On harmonic approximation in the $\mathcal{C}^1$ norm. *Mat. Sb.* **181** (10) (1990), 1341–1365; English transl. *Math. USSR-Sb.* **71** (1) (1992), 183–207.

[PS] Peres, Y., and Solomyak, B., How likely is Buffon's needle to fall near a planar Cantor set? *Pacific J. Math.* **204** (2) (2002), 473–496.

[Pr] Prat, L., Potential theory of signed Riesz kernels: capacity and Hausdorff measures. *Internat. Math. Res. Notices* **2004** (19) (2004), 937–981.

[Sch] Schul, R., Characterization of Subsets of Rectifiable Curves in Hilbert Space and the Analyst's Traveling Salesman Problem. Ph. Thesis, Yale, 2005.

[Su] Suita, N., On subadditivity of analytic capacity for two continua. *Kodai Math. J.* **7** (1984), 73–75.

[To1] Tolsa, X., L^2-boundedness of the Cauchy integral operator for continuous measures. *Duke Math. J.* **98** (2) (1999), 269–304.

[To2] Tolsa, X., Principal values for the Cauchy integral and rectifiability. *Proc. Amer. Math. Soc.* **128** (7) (2000), 2111–2119.

[To3] Tolsa, X., Cotlar's inequality and existence of principal values for the Cauchy integral without the doubling condition. *J. Reine Angew. Math.* **502** (1998), 199–235.

[To4] Tolsa, X., On the analytic capacity γ_+. *Indiana Univ. Math. J.* **51** (2) (2002), 317–344.

[To5] Tolsa, X., Painlevé's problem and the semiadditivity of analytic capacity. *Acta Math.* **190** (1) (2003), 105–149.

[To6] Tolsa, X., The semiadditivity of continuous analytic capacity and the inner boundary conjecture. *Amer. J. Math.* **126** (2004), 523–567.

[To7] Tolsa, X., Bilipschitz maps, analytic capacity, and the Cauchy integral. *Ann. of Math.* **162** (3) (2005), 1241–1302.

[To8] Tolsa, X., L^2 boundedness of the Cauchy transform implies L^2 boundedness of all Calderón-Zygmund operators associated to odd kernels. *Publ. Mat.* **48** (2) (2004), 445–479.

[To9] Tolsa, X., Finite curvature of arc length measure implies rectifiability: a new proof. *Indiana Univ. Math. J.* **54** (4) (2005), 1075–1105.

[To10] X. Tolsa, Growth estimates for Cauchy integrals of measures and rectifiability. *Geom. Funct. Anal.*, to appear.

[TVe] Tolsa, X., and Verdera, J., May the Cauchy transform of a non-trivial finite measure vanish on the support of the measure? *Ann. Acad. Sci. Fenn. Math.*, to appear.

[Uy] Nguyen Xuan Uy, Removable sets of analytic functions satisfying a Lipschitz condition. *Ark. Mat.* **17** (1979), 19–27.

[Ve1] Verdera, J., A weak type inequality for Cauchy transforms of finite measures. *Publ. Mat.* **36** (1992), 1029–1034.

[Ve2] Verdera, J., Removability, capacity and approximation. In *Complex Potential Theory* (Montreal, PQ, 1993), NATO Adv. Sci. Int. Ser. C Math. Phys. Sci. 439, Kluwer Academic Publ., Dordrecht 1994, 419–473.

[Ve3] Verdera, J., On the $T(1)$-theorem for the Cauchy integral. *Ark. Mat.* **38** (2000), 183–199.

[Vi1] Vitushkin, A. G., Example of a set of positive length but of zero analytic capacity. *Dokl. Akad. Nauk SSSR* **127** (1959), 246–249 (Russian).

[Vi2] Vitushkin, A. G., Estimate of the Cauchy integral. *Mat. Sb.* **71** (113) (1966), 515–534 (Russian).

[Vi3] Vitushkin, A. G., The analytic capacity of sets in problems of approximation theory. *Uspekhi Mat. Nauk.* **22** (6) (1967), 141–199; English transl. *Russian Math. Surveys* **22** (1967), 139–200.

[Vo] Volberg, A., *Calderón-Zygmund capacities and operators on nonhomogeneous spaces*. CBMS Regional Conf. Ser. in Math. 100, Amer. Math. Soc., Providence, RI, 2003.

Institució Catalana de Recerca i Estudis Avançats (ICREA) and Departament de Matemàtiques, Universitat Autònoma de Barcelona, Spain

E-mail: xtolsa@mat.uab.es

The Brunn–Minkowski theorem and related geometric and functional inequalities

Franck Barthe*

Abstract. The Brunn–Minkowski inequality gives a lower bound of the Lebesgue measure of a sum-set in terms of the measures of the individual sets. It has played a crucial role in the theory of convex bodies. This topic has many interactions with isoperimetry or functional analysis. Our aim here is to report some recent aspects of these interactions involving optimal mass transport or the Heat equation. Among other things, we will present Brunn–Minkowski inequalities for flat sets, or in Gauss space, as well as local versions of the theorem which apply to the study of entropy production in the central limit theorem.

Mathematics Subject Classification (2000). Primary 39B62; Secondary 52A40, 26D15, 94A17.

Keywords. Brunn–Minkowski, transport, Heat equation, entropy.

1. Introduction

The Brunn–Minkowski theory studies the relations between addition of vectors and the volume of convex sets. Let us start with some notation. For $\lambda \in \mathbb{R}$ and A a subset of $\mathbb{R}^d$, one sets $\lambda A = \{\lambda a;\ a \in A\}$. The Minkowski sum of two sets $A, B \subset \mathbb{R}^d$ is by definition

$$A + B := \{a + b;\ (a, b) \in A \times B\}.$$

The Brunn–Minkowski inequality gives a lower bound on the volume of a sum-set.

Theorem 1.1. *Let A, B be non-empty compact subsets of $\mathbb{R}^d$, then*

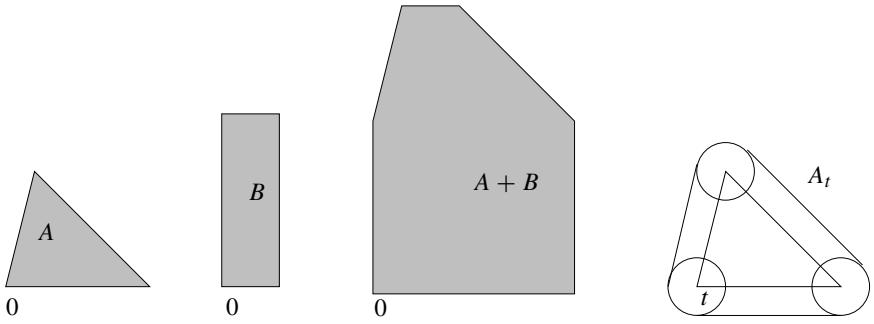
$$\mathrm{Vol}_d(A + B)^{\frac{1}{d}} \geq \mathrm{Vol}_d(A)^{\frac{1}{d}} + \mathrm{Vol}_d(B)^{\frac{1}{d}}. \tag{1}$$

If A, B are convex homothetic sets, there is equality. Brunn discovered this result in 1887 for A, B convex in dimension at most 3. Minkowski proved the inequality for convex sets in arbitrary dimension and realized the importance of the statement. Indeed, it could be combined with a former result by Steiner, which calculated the volume of the t-enlargement of a convex compact set $A \subset \mathbb{R}^3$ defined for $t > 0$ by

$$A_t = \{x \in \mathbb{R}^3;\ \text{there exists } y \in A \text{ such that } |x - y| \leq t\}.$$

*Research was supported in part by the European Network PHD, MCRN-511953.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006



Note that $A_t = A + tB^3$ where B^d denotes the Euclidean unit ball of $\mathbb{R}^d$. The Steiner formula asserts that for $t > 0$

$$\mathrm{Vol}_3(A + tB^3) = \mathrm{Vol}_3(K) + tS(A) + 2\pi t^2 W(A) + \frac{4}{3}\pi t^3,$$

where $S(A)$ is the surface area of A and $W(A)$ is its mean width (the average on unit vectors u of the width of the minimal slab orthogonal to u and containing A). The Brunn–Minkowski theorem provides relations between the above coefficients. Indeed, it implies that

$$\mathrm{Vol}_3(A + tB^3) \geq \left(\mathrm{Vol}_3(A)^{\frac{1}{3}} + t\mathrm{Vol}_3(B^3)^{\frac{1}{3}}\right)^3$$

with equality at $t = 0$. Comparing derivatives at zero gives

$$S(A) \geq 3\mathrm{Vol}_3(B^3)^{\frac{1}{3}}\mathrm{Vol}_3(A)^{\frac{2}{3}}.$$

This is the classical isoperimetric inequality; it means that among sets of given volume, balls have minimal surface area (the argument actually extends to non-convex sets). Another relation can be obtained by noting that the Brunn–Minkowski inequality shows that $\mathrm{Vol}_3(A + tB^3)^{\frac{1}{3}}$ is a concave function of $t \geq 0$ when A is convex.

Minkowski extended the Steiner formula as follows: for non-empty compact convex sets $K_1, \ldots, K_m \subset \mathbb{R}^d$ and $\lambda_1, \ldots, \lambda_m \geq 0$, the volume of $\lambda_1 K_1 + \cdots + \lambda_m K_m$ is a homogeneous polynomial of the form

$$\mathrm{Vol}_d(\lambda_1 K_1 + \cdots + \lambda_m K_m) = \sum_{i_1, \ldots, i_d = 1}^{m} \lambda_{i_1} \ldots \lambda_{i_d} V(K_{i_1}, \ldots, K_{i_d}).$$

Here and by definition $V(K_1, \ldots, K_d)$ is the mixed volume of d convex sets in $\mathbb{R}^d$. The theory of mixed volumes studies the properties of these quantities, their geometric interpretations and the inequalities among them. We refer to the book [53] for more on this topic. See also [52] where such volume estimates are used in the local theory of Banach spaces.

2. Functional extensions, functional tools

There exist several proofs of the Brunn–Minkowski theorem, see the surveys [34], [36] for details and precise references. However the most fruitful approach is probably the one based on the following functional version of the statement:

Theorem 2.1 (Prékopa–Leindler). *Let f, g, h be measurable non-negative functions on $\mathbb{R}^d$ and $\lambda \in [0,1]$. If for all x, y in $\mathbb{R}^d$,*

$$h(\lambda x + (1-\lambda)y) \geq f^\lambda(x) g^{1-\lambda}(y),$$

then $\int_{\mathbb{R}^d} h \geq \left(\int_{\mathbb{R}^d} f\right)^\lambda \left(\int_{\mathbb{R}^d} g\right)^{1-\lambda}$.

Remark 2.2. When applied to characteristic functions of sets, the above result provides a multiplicative version of the Brunn–Minkowski inequality, which is equivalent to the one we stated. The functional inequality can be written with an outer integral, as

$$\int_{\mathbb{R}^d}^* \sup_{\lambda x+(1-\lambda)y=z} f^\lambda(x) g^{1-\lambda}(y)\, dz \geq \left(\int_{\mathbb{R}^d} f\right)^\lambda \left(\int_{\mathbb{R}^d} g\right)^{1-\lambda}.$$

It appears as a reverse form of the classical inequality of Hölder which asserts that the right hand side in the latter inequality is at least $\int f^\lambda g^{1-\lambda}$.

An elementary proof of the above inequality appears in [52]. Here we sketch another proof. Its main idea is quite old and appears e.g. in [38]. It contains implicitly the idea of measure transportation which allowed recent developments.

Proof. We work in dimension 1, the general case follows by induction. By approximation arguments one may restrict to positive continuous f and g. One introduces functions $x, y\colon [0,1] \to \mathbb{R}$ satisfying for $t \in [0,1]$

$$\int_{-\infty}^{x(t)} f = t \int f; \quad \int_{-\infty}^{y(t)} g = t \int g. \tag{2}$$

Consequently for $t \in [0,1]$ it holds $x'(t) f(x(t)) = \int f$ and $y'(t) g(y(t)) = \int g$. One defines a third function z on $[0,1]$ by $z(t) = \lambda x(t) + (1-\lambda) y(t)$. Our three functions are strictly increasing. Comparing geometric mean and arithmetic mean yields $z'(t) \geq (x'(t))^\lambda (y'(t))^{1-\lambda}$. Finally we use z as a change of variables to evaluate the integral of h. Using the above relations

$$\begin{aligned}\int h &\geq \int_0^1 h(z(t)) z'(t)\, dt \\ &\geq \int_0^1 f^\lambda(x(t)) g^{1-\lambda}(y(t)) (x'(t))^\lambda (y'(t))^{1-\lambda}\, dt \\ &= \left(\int f\right)^\lambda \left(\int g\right)^{1-\lambda}.\end{aligned}$$

□

2.1. Measure transportation. If μ, ν are two probability measures on $\mathbb{R}^n$, one says that a map $T : \mathbb{R}^d \to \mathbb{R}^d$ transports μ to ν if $T\mu = \nu$, meaning $\nu(B) = \mu(T^{-1}(B))$ for every Borel set B.

In dimension 1, a canonical choice always exist when the first measure has no atoms: one can choose T non-decreasing. Note that the maps x, y of the previous proof are particular non-decreasing transporting maps. Indeed, if μ is the Lebesgue measure restricted to [0, 1] and $d\nu(t) = f(t)dt/\int f$ the first relation in (2) can be rewritten as follows

$$\mu((-\infty, t]) = \nu((-\infty, x(t)]).$$

Since x is increasing and onto, this is equivalent to $\nu(B) = \mu(T^{-1}(B))$ for $B = (-\infty, s]$, and this relation extends to the Borel σ-field.

In higher dimension a remarkable analogue is available due to the works of Brenier [22] and McCann [48].

Theorem 2.3. *Let μ, ν be probability measures on $\mathbb{R}^d$. Assume that whenever $B \subset \mathbb{R}^d$ is a Borel set with Hausdorff dimension $d-1$ one has $\mu(B) = 0$. Then there exists a convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that the map $T = \nabla\Phi$ (defined a.e.) satisfies $\nu = T\mu$. The map T is uniquely determined almost everywhere too.*

If μ and ν have second moments, then among all maps S with $S\mu = \nu$, T minimizes the quadratic transportation cost

$$\int_{\mathbb{R}^d} |x - S(x)|^2 d\mu(x).$$

As recalled in the second part of the theorem, this monotone transport T is related to the theory of optimal transportation, which looks for the best way to ship some amount of material from a configuration to another one. We refer to the book [59] for more on this fascinating topic.

If we consider measures with densities ρ_μ, ρ_ν with respect to Lebesgue's measure, then Φ is a generalized solution for the Monge–Ampère equation

$$\rho_\mu(x) = \rho_\nu(\nabla\Phi(x)) \det\big(\mathrm{Hess}\Phi(x)\big).$$

Weak and strong regularity theory for this equation were developed respectively by McCann [47] and Caffarelli [23]. McCann also introduced the following interpolation between the measures μ and ν: $((1-t)I + tT)\mu = \nabla(x \mapsto (1-t)|x|^2/2 + t\Phi(x))\mu$ for $t \in [0, 1]$. He found applications to equilibrium states (and also to a proof of the Brunn–Minkowski inequality). Optimal transport allows to interpolate between general densities. However it has more structure when Gaussian measures are involved, as Caffarelli proved:

Theorem 2.4 ([24]). *Let Q be a positive definite quadratic form on $\mathbb{R}^d$ let $d\mu(x) = e^{-Q(x)}dx/Z$ be the corresponding Gaussian probability measure. Let $d\nu = \rho\, d\mu$ be another probability measure with log-concave density ρ with respect to μ (i.e. $\rho(\lambda x + (1_\lambda)y) \geq \rho(x)^\lambda \rho(y)^{1-\lambda}$ for $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$). Then the monotone*

transportation map T such that $T\mu = \nu$ is a contraction for the canonical Euclidean distance.

2.2. Heat equation. Let $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ be the Laplace operator in $\mathbb{R}^d$. Following the probabilistic normalization we define the heat semigroup as $P_t = e^{t\Delta/2}$. More precisely, for a function f on $\mathbb{R}^d$ the function $u(t,x) = P_t f(x)$ solves the equation $\partial_t u = \frac{1}{2}\Delta u$ on $\mathbb{R}^+ \times \mathbb{R}^d$ with initial condition $u(0,\cdot) = f$. When f has three bounded derivatives, one has

$$P_t f(x) = \int_{\mathbb{R}^d} f(z) e^{-\frac{(z-x)^2}{2t}} \frac{dz}{(2\pi t)^{d/2}} \quad \text{for } t > 0.$$

C. Borell discovered that the heat flow preserves in some sense the hypothesis of the Prékopa–Leindler inequality. More precisely, given $\lambda \in (0,1)$ and three sufficiently regular non-negative functions $f, g, h\colon \mathbb{R}^d \to \mathbb{R}^+$ satisfying

$$h(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda},$$

he proved [18] that for all $t > 0$ and all x, y the following is true

$$P_t h(\lambda x + (1-\lambda)y) \geq P_t f(x)^\lambda P_t g(y)^{1-\lambda}.$$

The Prékopa–Leindler inequality is obtained in the limit $t \to +\infty$ since

$$P_t f(x) \sim_{t\to+\infty} (2\pi t)^{-d/2} \Big(\int f(y)dy \Big).$$

Borell's method was recently applied with success to derive other important Brunn–Minkowski type results. We will describe them in the next sections.

Remark 2.5. It has been known for many years that the heat equation is a powerful tool to prove functional inequalities of geometric flavor. In particular Bakry and Emery developed a general framework for deriving logarithmic Sobolev inequalities (which ensure Gaussian concentration of measure), or Sobolev type inequalities by semigroup techniques. More recently Bakry and Ledoux where able to prove Bobkov's functional form of the Gaussian isoperimetric inequality along these lines. It was also observed that the Brunn–Minkowski inequality implies various types of isoperimetric inequality. So morally, the use of the heat equation for Brunn–Minkowski type inequalities is not a complete surprise. The interested reader will find details in [3], [44], [43]. Recently the transportation method also allowed to derive concentration estimates and Sobolev type inequalities, see e.g. [59], [32].

2.3. Riemannian manifolds. McCann [49] has solved the optimal transport problem on a Riemannian manifold when the transportation cost is the square of the geodesic distance. This provides a natural generalization of the monotone map, and allowed remarkable extensions of the Prékopa–Leindler inequality by Cordero-Erausquin, McCann and Schmuckenschläger [31], [30]. The following statement is valid under a curvature assumption in the spirit of Bakry–Emery

Theorem 2.6 ([30]). *Let (M, g) be a Riemannian manifold, and let μ be a measure on M with density e^{-V} with respect to the volume measure. Assume that for $\rho \in \mathbb{R}$, the Ricci curvature and the Hessian of V satisfy*

$$\mathrm{Hess}_x V + \mathrm{Ric}_x \geq \rho g$$

for all $x \in M$. Let $\lambda \in [0,1]$ and $f, g, h : M \to \mathbb{R}^+$ such that for all $x, y \in M$ and all z such that $d(x,z) = (1-\lambda)d(x,y)$ and $d(z,y) = \lambda d(x,y)$ one has

$$h(z) \geq e^{-\rho d^2(x,y)/2} f^{\lambda}(x) g^{1-\lambda}(y),$$

then one gets: $\int_M h\, d\mu \geq \left(\int_M f\, d\mu\right)^{\lambda} \left(\int_M g\, d\mu\right)^{1-\lambda}$.

The condition on the intermediate point z (involving geodesic distances) simply means that z is a geodesic barycenter of x, y with weights λ, $1-\lambda$. Unlike in Euclidean spaces, there might be many of them.

3. Multilinear inequalities

The Brascamp–Lieb [21] inequalities are a powerful extension of Hölder's inequality. Their original motivation was the calculation of the best constant in Young's convolution inequality. Their most general form was established by Lieb. The setting of the theorem is the following. For $1 \leq i \leq m$, one considers linear surjective maps $B_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ and numbers $c_i \in [0,1]$.

Theorem 3.1 (Lieb [45]). *The best constant $K \in [0, +\infty]$ such that the inequality*

$$\int_{\mathbb{R}^n} \prod_{i=1}^m f_i(B_i x)^{c_i}\, dx \leq K \prod_{i=1}^m \left(\int_{\mathbb{R}} f_i\right)^{c_i}$$

holds for all integrable functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^+$ can be computed by considering only centered Gaussian functions $f_i(x) = \exp(-\langle A_i x, x\rangle)$ where the A_i's are symmetric positive definite matrices of size n_i.

Homogeneity shows that the constant may be finite only when $\sum_{i=1}^m c_i n_i = n$. This condition is assumed in the following. Since Gaussian integral may be computed, one gets $K = D^{-1/2}$ where

$$D = \inf_{A_i > 0} \frac{\det\left(\sum_{i=1}^m c_i B_i^* A_i B_i\right)}{\prod_{i=1}^m \det(A_i)^{c_i}}. \tag{3}$$

Here B_i^* denotes the adjoint of B_i. The proofs of Brascamp–Lieb and Lieb relied partially on tensorization arguments in higher dimension. We gave another proof using the monotone transport when proving an extension of the Prékopa–Leindler

inequality conjectured by K. Ball. The inequality is a reverse form of the Brascamp–Lieb inequality. The argument of proof is sophistication of the one given in the previous section, and gives both inequalities at a time. It uses the fact that the Jacobian matrices of monotone transport are symmetric positive matrices. This matches exactly the quantity appearing in the calculation of the Gaussian constant (3). The statement is

Theorem 3.2 ([10]). *The best constant $L \geq 0$ such that for all integrable functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^+$ one has*

$$\int_{\mathbb{R}^n} \sup_{\sum c_i B_i^* x_i = x;\ x_i \in \mathbb{R}^{n_i}} \prod_{i=1}^m f_i(x_i)^{c_i}\, dx \geq L \prod_{i=1}^m \left(\int_{\mathbb{R}^{n_i}} f_i \right)^{c_i},$$

can be computed on centered Gaussian functions, and $L = \sqrt{D}$.

Our motivation for studying these inequalities came from convex geometry. Ball first understood the relevance of the Brascamp–Lieb inequality for this topic. In the case $n_i = 1$, $B_i(x) = \langle x, u_i \rangle$ where u_i are unit vectors in $\mathbb{R}^n$ with the additional condition

$$\mathrm{Id}_{\mathbb{R}^n} = \sum_{i=1}^m c_i P_{u_i}$$

(P_u is the orthogonal projection onto the line spanned by u) he showed that $D = 1$. So for non-negative functions on $\mathbb{R}$ one has

$$\int_{\mathbb{R}^n} \prod_{i=1}^m f_i^{c_i}(\langle x, u_i \rangle) dx \leq \prod_{i=1}^m \left(\int_{\mathbb{R}} f_i \right)^{c_i}.$$

Applied to characteristic functions of intervals this inequality gives an upper bound on the volume of an intersection of slabs. This was one of the crucial ingredients in Ball's exact estimates on slices of the cubes or on the volume ratios of convex bodies, see [4] for details. The reverse Brascamp–Lieb inequality allows to estimate from below the volumes of convex hull and of sums of flat sets. For example, we obtain the following extension of the Brunn–Minkowski inequality

Theorem 3.3. *Let $(E_i)_{i=1}^m$ are vector-subspaces of $\mathbb{R}^n$ and $c_i \in (0, 1]$ be such that $\mathrm{Id}_{\mathbb{R}^n} = \sum_{i=1}^m c_i P_{E_i}$. Set $n_i = \dim(E_i)$. If $K_i \subset E_i$ then*

$$\mathrm{Vol}_n\Big(\sum_{i=1}^m c_i K_i \Big) \geq \prod_{i=1}^m \mathrm{Vol}_{n_i}(K_i)^{c_i}.$$

It was recently understood that the Brascamp–Lieb inequalities can be derived using the heat equation. This new approach is due to Carlen, Lieb and Loss [26] for functions of one variable and was developed to full generality by Bennett, Carbery,

Christ and Tao [15]. One advantage is that it allows a better description of equality cases. However, contrary to the mass transport approach, the heat equation method requires to know in advance which Gaussian functions are best, and to find a way around when there is no best Gaussian function. It was necessary to understand more precisely the Gaussian optimization problem summed up in Equation (3) and to understand when the constant D is positive (this corresponds to a non-trivial inequality) and when it is achieved (i.e. when a Gaussian maximizer exists). We present answer to the first question, which is of independent interest. The case of functions of one variable has a more explicit solution:

Theorem 3.4 ([10], [26]). *Let $(u_i)_{i\le m}$ be non-zero vectors in $\mathbb{R}^n$. There exists a finite constant K such that for all integrable functions $f_i : \mathbb{R} \to \mathbb{R}^+$,*

$$\int_{\mathbb{R}^n} \prod_{i=1}^m f_i(\langle x, u_i\rangle)^{c_i} dx \le K \prod_{i=1}^m \left(\int_{\mathbb{R}} f_i\right)^{c_i}$$

if and only if $c = (c_1, \dots, c_m)$ belongs to the set

$$\begin{aligned}\mathcal{C} &= \mathrm{conv}\{\mathbf{1}_I;\ I \subset \{1,\dots,m\} \text{ and } (u_i)_{i\in I} \text{ is a basis}\}\\ &= \Big\{c \in \mathbb{R}^m_+;\ \textstyle\sum_{i=1}^m c_i n_i = n \text{ and for all } S \subset \{1,\dots,m\},\\ &\qquad\qquad \textstyle\sum_{i\in S} c_i \le \dim(\mathrm{Span}(u_i, i \in S))\Big\}.\end{aligned}$$

Here $\mathbf{1}_I$ is a vector in $\mathbb{R}^m$ whose ith coordinate is 1 if $i \in I$ and 0 otherwise.

In the general case, only a description by facets of the set of exponents leading to a finite constant (*domain of finiteness*) is available

Theorem 3.5 ([16], [15]). *There exist $K < +\infty$ such that for all $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^+$*

$$\int_{\mathbb{R}^n} \prod_{i=1}^m (f_i \circ B_i)^{c_i} \le K \prod_{i=1}^m \left(\int_{\mathbb{R}} f_i\right)^{c_i}$$

if and only if for all i, $c_i \ge 0$, $\sum_{i=1}^m c_i n_i = n$ and for all vector subspaces $V \subset \mathbb{R}^n$ it holds

$$\dim V \le \sum_{i=1}^m c_i \dim(B_i V).$$

Remark 3.6. In the interior of the domain of finiteness, Gaussian maximizers exist and the inequality is equivalent to the multidimensional version of Ball's form (also called the geometric form) of the Brascamp–Lieb inequality: if vector subspaces E_i and numbers $c_i \in (0, 1]$ are such that $\mathrm{Id}_{\mathbb{R}^n} = \sum_{i=1}^m c_i P_{E_i}$ then for non-negative functions $f_i : E_i \to \mathbb{R}^+$, one has

$$\int_{\mathbb{R}^n} \prod_{i=1}^m (f_i \circ P_{E_i})^{c_i} \le \prod_{i=1}^m \left(\int_{E_i} f_i\right)^{c_i}.$$

This is proved using the heat semigroup, showing that $t \mapsto \int \prod_{i=1}^m ((P_t f_i) \circ P_{E_i})^{c_i}$ is non-decreasing and interpolates between the two terms of the above inequality. On the boundary of the finiteness domain, a factorization argument allows to reduce the dimension and conclude by induction.

Remark 3.7. The reverse Brascamp–Lieb inequality can be proved by the heat flow too, along the lines of Borell's argument for the Prékopa–Leindler inequality. This is written in [12] for functions of one variables and the geometric form. However this easily extends as well as the others steps of the proof.

But the heat equation approach does not only provide us with new proofs. It allows remarkable extensions of the results together with new applications. We refer to [15] for inequalities restricted to special classes of functions. Carlen, Lieb and Loss where able to prove similar inequalities in other spaces as the sphere [26] and the symmetric group [25]. The spherical inequality was motivated by the study of a system of n particles in one dimension, preserving total kinetic energy. Hence their n speeds form a vector in the Euclidean sphere $S^{n-1} \subset \mathbb{R}^n$. In order to know how the information on an individual particle influences the one of the whole system, they established the following: for $f_i : [-1, 1] \to \mathbb{R}^+$,

$$\int_{S^{n-1}} \prod_{i=1}^m f_i(x_i)\, d\sigma(x) \leq \prod_{i=1}^m \left(\int_{S^{n-1}} f_i(x_i)^2 d\sigma(x) \right)^{\frac{1}{2}},$$

where σ is the uniform probability measure on the sphere. The surprise here is the 2-norm, which is best possible and in particular does not disappear when $n \to +\infty$. In [14] this is extended to general decompositions of the identity $\mathrm{Id}_{\mathbb{R}^n} = \sum_{i=1}^m c_i P_{E_i}$, where for functions $f_i : E_i \to \mathbb{R}^+$ it holds

$$\int_{S^{n-1}} \prod_{i=1}^m f_i(P_{E_i} x)^{c_i/2} d\sigma(x) \leq \prod_{i=1}^m \left(\int_{S^{n-1}} f_i(P_{E_i} x)\, d\sigma(x) \right)^{c_i/2}.$$

This allows to consider particle systems in d dimension, and also with fixed momentum for example. However the exponents $c_i/2$ may not be best possible in this generality. In the work [13] a general framework of commuting Markov generator is developed to deal with these inequalities in general settings. Also the geometric meaning of the best exponents is better understood in continuous settings, and several new examples are provided.

4. Geometry in Gauss space

Let us denote by γ_d the standard Gaussian probability measure on $\mathbb{R}^d$ with density with respect to Lebesgue's measure given by $\rho(x) = (2\pi)^{-d/2} \exp(-|x|^2/2)$, $x \in \mathbb{R}^d$. There is no need to emphasize its importance, and it is natural and useful to have

Brunn–Minkowski type inequalities for γ_d. Applying the Prékopa–Leindler theorem to $f = \rho \mathbf{1}_A$, $g = \rho \mathbf{1}_B$, where $A, B \subset \mathbb{R}^d$ and using the log-concavity of ρ yields

$$\gamma_d^*(\lambda A + (1-\lambda)B) \geq \gamma_d(A)^\lambda \gamma_d(B)^{1-\lambda}. \tag{4}$$

This inequality is not sharp. An optimal version was proved by Ehrhard [35] for convex sets, using a symmetrization procedure. Latała [40] showed next that one a the sets may be non-convex and recently Borell [19] completely removed the convexity assumption. His approach is functional and uses the Heat equation. The most general version of his result is given below.

Theorem 4.1 ([20]). *Let $\lambda, \mu \geq 0$ with $\lambda + \mu \geq 1$ and $|\lambda - \mu| \leq 1$. Then for all measurable sets $A, B \subset \mathbb{R}^d$ the following holds:*

$$\Phi^{-1}(\gamma_d^*(\lambda A + \mu B)) \geq \lambda \Phi^{-1}(\gamma_d(A)) + \mu \Phi^{-1}(\gamma_d(B)),$$

where Φ is the distribution function of γ_1, defined by $\Phi(t) = \int_{-\infty}^t e^{-u^2/2} du/\sqrt{2\pi}$ for $t \in \mathbb{R}$.

Here $\Phi^{-1}\colon [-1,1] \to [-\infty, +\infty]$ is the reciprocal of Φ and by convention $+\infty - \infty = -\infty$. The inequality becomes an equality when A and B are parallel half-spaces. It recovers and unifies classical results on dilates and enlargements of convex sets:

Corollary 4.2 ([56]). *Let A be a convex set in R^d, and let $H \subset \mathbb{R}^d$ be a half-space with $\gamma_d(A) = \gamma_d(H)$. Then for all $r \geq 1$,*

$$\gamma_d(rA) \geq \gamma_d(rH).$$

This is reversed when $r \in (0,1]$.

Corollary 4.3 ([56], [17]). *Let $A \subset R^d$ be measurable, and $H \subset \mathbb{R}^d$ be a half-space such that $\gamma_d(A) = \gamma_d(H)$. Then for all $r \geq 0$,*

$$\gamma_d(A + rB^d) \geq \gamma_d(H + rB^d).$$

The latter statement is the sharp Gaussian isoperimetric inequality. It implies among others the concentration phenomenon (see e.g. [44]). It asserts that every L-Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ is close to its median M with high probability:

$$\gamma_d\left(|f - M| \geq t\right) \leq 2e^{-t^2/(2L^2)}.$$

This fact is of fundamental importance in particular in the geometry of Banach spaces [50], [52], where it is often applied to norms. This is one of the motivations for studying improvements of the above results for symmetric convex sets. Little is known in this direction. Latała–Oleszkiewicz [41]] established a sharp analogue of Corollary 4.2 for symmetric convex sets. Let us mention another result of similar flavor. Caffarelli's Theorem 2.4 on monotone transportation nicely enters its proof.

Theorem 4.4 (Cordero-Erausquin–Fradelizi–Maurey [29]). *Let $A \subset \mathbb{R}^d$ be an origin-symmetric convex set. Then the function $t \mapsto \log(\gamma_d(tA))$ is concave on $(0, +\infty)$.*

In other words for $\lambda \in (0,1)$ and $s, t > 0$, $\gamma_d(s^\lambda t^{1-\lambda} A) \geq \gamma_d(sA)^\lambda \gamma_d(tA)^{1-\lambda}$. As $s^\lambda t^{1-\lambda} A \subset (\lambda s + (1-\lambda)t)A$, this is an improvement on what could be obtained before from the Prékopa–Leindler inequality

$$\gamma_d((\lambda s + (1-\lambda)t)A) \geq \gamma_d(sA)^\lambda \gamma_d(tA)^{1-\lambda}.$$

This fact and the isoperimetric inequality were used to derive the following improvement of the Gaussian concentration of norm, as conjectured by Vershynin.

Theorem 4.5 (Latała–Oleszkiewicz [42]). *Let G be a standard Gaussian vector in $(\mathbb{R}^d, \|\cdot\|)$. Let M be a median of $\|G\|$ and $\sigma^2 = \sup_{\|f\|_* \leq 1} Ef^2(G)$. Then for all $t \in (0,1]$ one has*

$$P(\|G\| \leq tM) \leq \frac{1}{2}(2t)^{M^2/(4\sigma^2)}.$$

Remark 4.6. Finally let us point out that Caffarelli's contraction Theorem 2.4 also played a crucial role in the recent progress towards the Gaussian correlation conjecture which predicts that every two origin symmetric convex sets $A, B \subset \mathbb{R}^d$ satisfy $\gamma_d(A \cap B) \geq \gamma_d(A)\gamma_d(B)$. See [28], [37].

5. Shannon entropy

Let X be a random variable with density $f : \mathbb{R} \to [0, \infty)$ and, to fix ideas, such that $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$. Its Shannon entropy is by definition

$$\mathrm{Ent}(X) = -\int_{\mathbb{R}} f \log f.$$

This fundamental notion of information theory also plays a crucial role in the study of return to equilibrium of many random systems. In this section we are interested in entropic aspects of the Central Limit Theorem (CLT). A new approach to entropy estimates was developed, which was formally inspired by a local version of the Brunn–Minkowski theorem.

Among variables of given variance, Gaussian variables are known to maximize entropy. In other words, if G is a standard Gaussian variable with density given by $g(t) = (2\pi)^{-1} \exp(-|t|^2/2)$, $t \in \mathbb{R}$, it holds that

$$\mathrm{Ent}(X) \leq \mathrm{Ent}(G).$$

Moreover, the difference between these two entropies is a strong distance between the laws of X and G. Indeed, the Pinsker–Csiszar–Kullback inequality [51], [33], [39] asserts that it dominates the square of the total variation distance. If Y, Z are

independent random variables and $\lambda \in (0,1)$, the Shannon–Stam inequality [54], [55] asserts that

$$\lambda \mathrm{Ent}(Y) + (1-\lambda)\mathrm{Ent}(Z) \leq \mathrm{Ent}(\sqrt{\lambda}\, Y + \sqrt{1-\lambda}\, Z). \tag{5}$$

In particular if $(X_i)_{i\geq 1}$ are independent copies of X one gets that

$$\mathrm{Ent}(X_1) \leq \mathrm{Ent}\left(\frac{X_1+X_2}{\sqrt{2}}\right),$$

and by iteration that

$$\mathrm{Ent}\left(\frac{1}{\sqrt{2^k}}\sum_1^{2^k} X_i\right)$$

is non-decreasing k (and bounded from above by the entropy of the standard Gaussian variable). Linnik [46] was the first to prove the CLT using entropy. Next Barron [9] established the CLT with entropic convergence. Obtaining rates for the convergence of the entropy requires to improve on the Shannon–Stam inequality (5). Carlen and Soffer [27] obtained non-explicit results in this direction.

In [5] Ball, Naor and the author developed a new technique to estimate entropy production. It is based on a new representation of the Fisher information of a marginal. Recall that the Fisher information of a variable X with density f is defined as $I(X) = I(f) := \int (f')^2/f$. It corresponds to the derivative of entropy along the Ornstein–Uhlenbeck semigroup: let G be a standard Gaussian variable independent of X; set $X_t := \sqrt{e^{-t}}\, X + \sqrt{1-e^{-t}}\, G$ and let f_t denote its density. Then

$$\mathrm{Ent}(G) - \mathrm{Ent}(X) = \int_0^\infty (I(X_t) - I(G))\, dt.$$

This classical relation allows to integrate linear inequalities on I in order to derive entropic estimates. The Fisher information representation was inspired by the Brunn–Minkowski theorem, as explained in the following section.

Remark 5.1. There was already a nice connection with Brunn–Minkowski theory. Indeed an equivalent form of the Shannon–Stam inequality (5) known as the entropy power inequality asserts that for independent random variables Y, Z one has

$$e^{2\mathrm{Ent}(Y+Z)} \geq e^{2\mathrm{Ent}(Y)} + e^{2\mathrm{Ent}(Z)}.$$

The similarity with the Brunn–Minkowski theorem was noted early and it was supported by the interpretation of Shannon entropy in terms of volumes of typical sets of values of independent copies of a variable. This analogy as well as the occurrence of the number 2 was explained by Szarek and Voiculescu [57], [58], who derived the entropy power inequality from a *restricted* Brunn–Minkowski inequality. Mass transport allows to establish a functional version of the latter, see [11].

5.1. A local version of the Brunn–Minkowski theorem. Consider a probability density $w(x, y)$ on $\mathbb{R}^2$ together with the density of its first marginal.

$$h(x) = \int w(x, y)\, dy.$$

Under appropriate regularity and integrability assumptions, the Fisher information of the marginal is expressed in terms of $(\log h)''$:

$$I(h) = \int \frac{(h')^2}{h} = \int h'(\log h)' = -\int h(\log h)''.$$

A direct consequence of the Prékopa–Leindler theorem is that $\log h$ is concave when $\log w$ is (this fact is actually formally equivalent to Brunn–Minkowski for convex sets). In [5] this is explained in terms of second derivatives. The direct calculation is not conclusive and has to be rearranged as

$$h(x)(-\log h)''(x) = \int_{\mathbb{R}} w(x, y)\left[(\partial_y \rho)^2(x, y) + \mathrm{D}^2(-\log w)_{(x,y)} \cdot \begin{pmatrix} 1 \\ \rho(x, y) \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \rho(x, y) \end{pmatrix}\right] dy$$

where ρ is defined by

$$\rho(x, y) = \frac{1}{w(x, y)} \left(\frac{h'(x)}{h(x)} \int_{-\infty}^{y} w(x, v)\, dv - \int_{-\infty}^{y} \partial_x w(x, v)\, dv \right).$$

One easily reads on the first formula that $\mathrm{D}^2(\log w) \leq 0$ implies $(\log h)'' \leq 0$. Actually the function $y \mapsto \rho(x, y)$ described above minimizes the term on the right. Hence we get the following representation

$$h(x)(-\log h)''(x) = \inf_{p:\, \mathbb{R} \to \mathbb{R}} \int_{\mathbb{R}} w(x, y)\left[(p'(y))^2 + \mathrm{D}^2(-\log w)_{(x,y)} \cdot \begin{pmatrix} 1 \\ p(y) \end{pmatrix} \cdot \begin{pmatrix} 1 \\ p(y) \end{pmatrix}\right] dy.$$

Integrating with respect to x yields an expression of the Fisher information of the marginal.

5.2. Variational expressions for the Fisher information. In subsequent papers with Artstein, Ball and Naor [2], [1], more intrinsic formulations are given: let $w: \mathbb{R}^n \to \mathbb{R}^+$ be a probability density and let $e \in S^{n-1}$ be a unit vector. One considers the marginal obtained by projection onto $\mathbb{R}e$:

$$h(t) = \int_{te+e^\perp} w.$$

Its Fisher information can be expressed as an infimum in the following ways:

$$\begin{aligned} I(h) &= \inf_k \int_{\mathbb{R}^n} wk^2 \\ &= \inf_q \int_{\mathbb{R}^n} w \left(\frac{\operatorname{div}(wq)}{w} \right)^2 \\ &= \inf_q \int_{\mathbb{R}^n} w \left[\operatorname{Tr}(Dq)^2 + D^2(-\log w).q.q \right] \end{aligned}$$

where the first infimum is over functions $k : \mathbb{R}^n \to \mathbb{R}$ such that for all $t \in \mathbb{R}$ it holds $\int_{te+e^\perp} \partial_e w = \int_{te+e^\perp} wk$. The last two infima range over applications $q : \mathbb{R}^n \to \mathbb{R}^n$ such that for all $x \in \mathbb{R}^n$ one has $\langle q(x), e \rangle = 1$ (we have omitted here a few regularity conditions).

These formulas are convenient tools to estimate the Fisher information of $(X_1 + \cdots + X_k)/\sqrt{k}$ which is a marginal of $(X_1, \ldots, X_k)$. Applications are presented next.

5.3. The monotonicity of entropy in the CLT. The paper [2] answers an old conjecture of Shannon by showing that if $(X_n)_{n \geq 1}$ are independent copies of square integrable a random variable X with finite entropy then the sequence

$$e_k := \operatorname{Ent}\left(\frac{X_1 + \cdots + X_k}{\sqrt{k}} \right)$$

is non-decreasing. The classical Shannon–Stam inequality gives $e_k \leq e_{2k}$, but $e_k \leq e_{k+1}$ is much harder. This is deduced from a similar fact for Fisher information, which is proved using the infimum representation: the best test function for k variables is used to build a suitable test functions for $k + 1$. A corresponding version of the entropy-power inequality is also proved

Theorem 5.2 ([2]). *Let $X_1, \ldots, X_{n+1}$ be independent square integrable random variables. Then*

$$\exp\Big[2\operatorname{Ent}\Big(\sum_{i=1}^{n+1} X_i\Big)\Big] \geq \frac{1}{n} \sum_{j=1}^{n+1} \exp\Big[2\operatorname{Ent}\Big(\sum_{i \neq j} X_i\Big)\Big].$$

5.4. Rate of convergence in the entropic CLT. In [5], [1] the rate of entropy production, when adding independent copies, is studied under a spectral gap hypothesis. Recall that a random variable X has a spectral gap (or satisfies a Poincaré inequality) if there exists $c > 0$ such that every smooth function $s : \mathbb{R} \to \mathbb{R}$ verifies

$$c(E(s(X)^2) - (Es(X))^2) \leq E(s'(X)^2). \tag{6}$$

The strategy of proof was to choose specific functions in the variational formula for the Fisher information. Barron and Johnson where able to recover this result by a different method [8].

Theorem 5.3 ([1]). *Let G be a standard Gaussian random variable and let X be a random variable with variance* 1. *Assume that X satisfies a spectral gap inequality with constant $c > 0$. If $X_1, \dots, X_n$ are independent copies of X, denote as usual $S_n = (X_1 + \cdots + X_n)/\sqrt{n}$. Then*

$$\mathrm{Ent}(G) - \mathrm{Ent}(S_n) \leq \frac{1}{1 + \frac{c}{2}(n-1)} (\mathrm{Ent}(G) - \mathrm{Ent}(X)).$$

The rate $1/n$ is best possible. The spectral gap assumption is easy to decide on the real line. It is however rather strong, as it imply in particular exponential integrability. It is natural to try and replace this assumption with weaker moment conditions. A first quantitative result in this direction was obtained by Ball and Cordero-Erausquin:

Theorem 5.4 ([6]). *Let X be a symmetric random variable with $E(X^2) = 1$. Assume that it has finite Fisher information $I(X)$ and third moment $\tau = (E|X|^3)^{1/3}$. Then for $n \geq 1$*

$$\mathrm{Ent}(G) - \mathrm{Ent}(S_n) \leq \frac{c\sqrt{\tau}\, I(X)^{3/2}}{n^{\alpha}},$$

where c, α are universal constants.

Remark 5.5. Entropy production in the case of Markov chains with spectral gap was recently understood [7].

References

[1] Artstein, S., Ball, K., Barthe, F., and Naor, A., On the rate of convergence in the entropic central limit theorem. *Probab. Theory Related Fields* **129** (2004), 381–390.

[2] Artstein, S., Ball, K., Barthe, F., and Naor, A., Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.* **17** (2004), 975–982.

[3] Bakry, D., L'hypercontractivité et son utilisation en théorie des semigroupes. In *Lectures on probability theory* (Saint-Flour, 1992), Lecture Notes in Math. 1581, Springer, Berlin 1994, 1–114.

[4] Ball, K., Convex geometry and functional analysis. In *Handbook of the geometry of Banach spaces* (ed. by W. B. Johnson et al.), Volume 1, Elsevier, Amsterdam 2001, 161–194.

[5] Ball, K., Barthe, F., and Naor, A., Entropy jumps in the presence of a spectral gap. *Duke Math. J.* **119** (1) (2003), 41–63.

[6] Ball, K., and Cordero-Erausquin, D., Convergence of information in the Central Limit Theorem under moment conditions. Preprint, 2005.

[7] Ball, K., Martin-Marquez, V., and Naor, A., Rapid entropy growth for Markov chains with a spectral gap. In preparation, 2005.

[8] Barron A., and Johnson, O., Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields* **129** (3) (2004), 391–409.

[9] Barron, A. R., Entropy and the central limit theorem. *Ann. Probab.* **14** (1986), 336–342.

[10] Barthe, F., On a reverse form of the Brascamp-Lieb inequality. *Invent. Math.* **134** (1998), 335–361.

[11] Barthe, F., Restricted Prékopa-Leindler inequality. *Pacific J. Math.* **189** (1999), 211–222.

[12] Barthe, F., and Cordero-Erausquin, D., Inverse Brascamp-Lieb inequalities along the Heat equation. In *Geometric Aspects of Functional Analysis* (ed. by V. D. Milman and G. Schechtman), Lecture Notes in Math. 1850, Springer-Verlag, Berlin 2004, 65–71.

[13] Barthe, F., Cordero-Erausquin, D., Ledoux, M., and Maurey, B., Semigroup proofs of Brascamp-Lieb inequalities. In preparation, 2006.

[14] Barthe, F., Cordero-Erausquin, D., and Maurey, B., Entropy of spherical marginals. *J. Math. Pures Appl.*, to appear.

[15] Bennett, J., Carbery, A., Christ, M., and Tao, T., The Brascamp-Lieb inequalities: finiteness, structure and extremals. Preprint, 2005.

[16] Bennett, J., Carbery, A., Christ, M., and Tao, T., Finite bounds for Hölder-Brascamp-Lieb multilinear inequalities. Preprint, 2005.

[17] Borell, C., The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30** (1975), 207–216.

[18] Borell, C., Diffusion equations and geometric inequalities. *Potential Anal.* **12** (1) (2000), 49–71.

[19] Borell, C., The Ehrhard inequality. *C. R. Math. Acad. Sci. Paris* **337** (10) (2003), 663–666.

[20] Borell, C., Minkowski sums in Gaussian analysis. In *Notes de l'école d'hiver "Probabilistic Methods in High Dimension Phenomena"* (Toulouse, 2005); http://www.lsp.ups-tlse.fr/Proba_Winter_School/.

[21] Brascamp, H. J., and Lieb, E. H., Best constants in Young's inequality, its converse and its generalization to more than three functions. *Adv. Math.* **20** (1976), 151–173.

[22] Brenier, Y., Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** (1991), 375–417.

[23] Caffarelli, L., The regularity of mappings with a convex potential. *J. Amer. Math. Soc.* **4** (1992), 99–104.

[24] Caffarelli, L. A., Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.* **214** (3) (2000), 547–563; Erratum ibid. **225** (2) (2002), 449–450.

[25] Carlen, E. A., Lieb, E. H., and Loss, M., An inequality of Hadamard type for permanents. Preprint.

[26] Carlen, E. A., Lieb, E. H., and Loss, M., A sharp analog of Young's inequality on S^N and related entropy inequalities. *J. Geom. Anal.* **14** (3) (2004), 487–520.

[27] Carlen, E. A., and Soffer, A., Entropy production by block variable summation and central limit theorem. *Comm. Math. Phys.* **140** (2) (1991), 339–371.

[28] Cordero-Erausquin, D., Applications of mass transport to Gaussian-type inequalities. *Arch. Rational Mech. Anal.* **161** (2002), 257–269.

[29] Cordero-Erausquin, D., Fradelizi, M., and Maurey, B., The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems. *J. Funct. Anal.* **214** (2) (2004), 410–427.

[30] Cordero-Erausquin, D., McCann, R. J., and Schmuckenschläger, M., Prékopa-Leindler type inequalities on riemannian manifolds, mass transport and Jacobi fields. Preprint.

[31] Cordero-Erausquin, D., McCann, R. J., and Schmuckenschläger, M., A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math.* **146** (2) (2001), 219–257.

[32] Cordero-Erausquin, D., Nazaret, B., and Villani, C., A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. *Adv. Math.* **182** (2) (2004), 307–332.

[33] Csiszar, I., Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* Ser. A **7** (1962), 137–157.

[34] Das-Gupta, S., Brunn-Minkowski inequality and its aftermath. *J. Multivariate Anal.* **10** (1980), 296–318.

[35] Ehrhard, A., Symétrisation dans l'espace de Gauss. *Math. Scand.* **53** (1983), 281–301.

[36] Gardner, R. J., The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.* (*N.S.*) **3** (2002), 355–405.

[37] Hargé, G., A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces. *Probab. Theory Related Fields* **130** (3) (2004), 415–440.

[38] Henstock, R., and Macbeath, A. M., On the measure of sum sets. (I) the theorems of Brunn, Minkowski and Lusternik. *Proc. London Math. Soc.* **3** (1953), 182–194.

[39] Kullback, S., A lower bound for discrimination information in terms of variation. *IEEE Trans. Inform. Theory* **4** (1967), 126–127.

[40] Latała, R., A note on the Ehrhard inequality. *Studia Math.* **118** (2) (1996), 169–174.

[41] Latała, R., and Oleszkiewicz, K., Gaussian measures of dilatations of convex symmetric sets. *Ann. Probab.* **27** (4) (1999), 1922–1938.

[42] Latała, R., and Oleszkiewicz, K., Small ball probability estimates in terms of width. *Studia Math.* **169** (3)(2005), 305–314.

[43] Ledoux, M., The geometry of Markov diffusion generators. *Ann. Fac. Sci. Toulouse Math.* (6) **9** (2) (2000), 305–366.

[44] Ledoux, M., *The concentration of measure phenomenon*. Math. Surveys Monogr. 89, Amer. Math. Soc., Providence, RI, 2001.

[45] Lieb, E. H., Gaussian kernels have only gaussian maximizers. *Invent. Math.* **102** (1990), 179–208.

[46] Linnik, Ju. V., An information theoretic proof of the central limit theorem with lindeberg conditions. *Theory Probab. Appl.* **4** (1959), 288–299.

[47] McCann, R. J., A Convexity Theory for Interacting Gases and Equilibrium Crystals. PhD thesis, Princeton University, 1994.

[48] McCann, R. J., Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80** (2) (1995), 309–323.

[49] McCann, R. J., Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.* **11** (3) (2001), 589–608.

[50] Milman, V., and Schechtman, G., *Asymptotic Theory of Finite Dimensional Normed Spaces*. Lecture Notes in Math. 1200, Springer-Verlag, Berlin 1986.

[51] Pinsker, M. Š., *Information and information stability of random variables and processes*. Holden-Day, San Francisco, 1964.

[52] Pisier, G., *The volume of convex bodies and Banach space geometry*. Cambridge Tracts in Math. 94, Cambridge University Press, Cambridge 1989.

[53] Schneider, R., *Convex bodies: the Brunn-Minkowski theory*. Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge 1993.

[54] Shannon, C. E., and Weaver, W., *The mathematical theory of communication*. University of Illinois Press, Urbana, IL, 1949.

[55] Stam, A. J., Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. and Control* **2** (1959), 101–112.

[56] Sudakov, V. N., and Tsirel'son, B. S., Extremal propreties of half-spaces for spherically invariant measures. *J. Soviet Math.* **9** (1978), 9–18; Russian original *Zap. Nauchn. Sem. Leningrad. Otdel. Math. Inst. Steklova.* **41** (1974) 14–24.

[57] Szarek, S., and Voiculescu, D., Volumes of restricted Minkowski sums and the free analogue of the entropy power inequality. *Comm. Math. Phys.* **178** (3) (1996), 563–570.

[58] Szarek, S., and Voiculescu, D., Shannon's entropy power inequality via restricted Minkowski sums. In *Geometric aspects of functional analysis*, Lecture Notes in Math. 1745, Springer-Verlag, Berlin 2000, 257–262.

[59] Villani, C., *Topics in optimal transportation*. Grad. Stud. Math. 58, Amer. Math. Soc., Providence, RI, 2003.

Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex 09, France
E-mail: barthe@math.ups-tlse.fr

Isomorphic and almost-isometric problems in high-dimensional convex geometry

Bo'az Klartag*

Abstract. The classical theorems of high-dimensional convex geometry exhibit a surprising level of regularity and order in arbitrary high-dimensional convex sets. These theorems are mainly concerned with the rough geometric features of general convex sets; the so-called "isomorphic" features. Recent results indicate that, perhaps, high-dimensional convex sets are also very regular on the almost-isometric scale. We review some related research directions in high-dimensional convex geometry, focusing in particular on the problem of geometric symmetrization.

Mathematics Subject Classification (2000). 52A20, 52A21, 46B07.

Keywords. Convex geometry, high dimension, concentration phenomenon, central limit theorem.

1. Introduction

We will begin by quoting a sample of two fundamental theorems from the asymptotic theory of finite-dimensional normed spaces. The first will be Dvoretzky's theorem (for proofs, credits and history see, e.g., [47], [58] and references therein). We work in $\mathbb{R}^n$, endowed with the standard Euclidean norm $|\cdot|$ and scalar product $\langle \cdot, \cdot \rangle$. A convex body in $\mathbb{R}^n$ is a compact, convex set with a non-empty interior.

Theorem 1.1 (Dvoretzky's theorem). *Let $K \subset \mathbb{R}^n$ be a convex body that is centrally-symmetric (i.e. $K = -K$) and let $0 < \varepsilon < 1$. Then there exist $r > 0$ and a subspace $F \subset \mathbb{R}^n$ with $\dim(F) > c\varepsilon^2 \log n$ such that*

$$(1-\varepsilon) r D_F \subset K \cap F \subset (1+\varepsilon) r D_F,$$

where $D_F = \{x \in F;\ |x| \leq 1\}$ is the Euclidean unit ball in the subspace F and $c > 0$ is a universal constant.

Theorem 1.1 reveals a basic property of centrally-symmetric convex sets in high dimension: They all contain almost-spherical sections of logarithmic dimension. The second theorem we quote is Milman's quotient of subspace theorem [53]. It presents an almost full-dimensional approximate ellipsoid that is "hidden" in a certain way within any convex body in high dimension.

*The author is a Clay Research Fellow. Partial support was also given by NSF grant #DMS-0456590.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

Theorem 1.2 (Milman's quotient of subspace theorem). *Let $K \subset \mathbb{R}^n$ be a centrally-symmetric convex body, and let $0 < \delta < \frac{1}{2}$. Then there exist subspaces $E \subset F \subset \mathbb{R}^n$ with $\dim(E) > \lceil (1-\delta)n \rceil$ and an ellipsoid $\mathcal{E} \subset E$ such that*

$$\frac{1}{c(\delta)} \mathcal{E} \subset \mathrm{Proj}_E(K \cap F) \subset c(\delta) \mathcal{E}.$$

Here Proj_E stands for the orthogonal projection operator onto E in $\mathbb{R}^n$ and $c(\delta) < c\frac{1}{\delta} \log \frac{1}{\delta}$, where $c > 0$ is a universal constant.

Asymptotic convex geometry is a discipline that emerged in the 1970*s* and 1980*s* from the geometric study of Banach spaces. It is also known by other names, such as the theory of high-dimensional normed spaces, asymptotic geometric analysis, etc. Theorem 1.1 and Theorem 1.2 are typical representatives of the achievements of asymptotic convex geometry. We refer the reader to, e.g., [31] for a more complete picture of this theory. Theorem 1.1, Theorem 1.2 and other results impose stringent regularity on the geometry of a general high-dimensional convex set (the central symmetry requirement is, in many cases, not entirely essential). An important feature of these results is their broad scope; a not-so-obvious fact that we learn from the asymptotic theory of convex geometry, is that there exist non-trivial, structural geometric statements that apply to all high-dimensional convex bodies.

The precise convexity is rarely used in this theory, and corresponding principles also hold under much weaker assumptions, such as quasi convexity. The focus is on the high dimension; the theory makes sense only when the dimension n is a very large number, tending to infinity. A protagonist in many proofs of high-dimensional results is the concentration of measure phenomenon, that is, the strong concentration inequalities that typical high-dimensional measures satisfy. This phenomenon and its applications were largely put forward by Milman, starting from his proof of Dvoretzky's theorem [52]. The concentration phenomenon forces regularity and simplicity on some apriori complicated objects such as a Lipshitz function on the sphere, and is one reason for the success of the high-dimensional theory (see, e.g., the review [54]).

A key characteristic of the theory is its "isomorphic" nature. That is, the scale in which convex bodies are viewed is such that two centrally-symmetric convex bodies $K, T \subset \mathbb{R}^n$ are considered to be "close enough" when

$$c_1 K \subset T \subset c_2 K \tag{1}$$

for $c_1, c_2 > 0$ being universal constants, independent of the dimension. In other words, the norms that have K and T as their unit balls, are uniformly isomorphic. This approach is most natural to functional analysis, the origin of the subject, and has led to an interesting and elegant theory. On the other hand, even some of the most basic questions of an "almost-isometric" nature in high dimension (as opposed to "isomorphic" nature) still remain unanswered. Let us present two such "almost-isometric" problems. The first is due to Bourgain [15], and will be discussed in more detail in Section 3. One of its many formulations reads as follows:

Question 1.1 (*The slicing problem*). Does there exist $c > 0$, such that for any dimension n and every convex body $K \subset \mathbb{R}^n$ of volume one, there exists at least one hyperplane section of K whose $(n-1)$-dimensional volume is larger than c?

The second question we would like to present, is the almost-isometric version of Dvoretzky's theorem.

Question 1.2. Fix a positive integer k. Do there exist $c(k), c'(k) > 0$ such that for any $0 < \varepsilon < 1$, $N = \left\lfloor c'(k) \left(\frac{1}{\varepsilon}\right)^{c(k)} \right\rfloor$ and for any centrally-symmetric convex body $K \subset \mathbb{R}^N$, one may find $r > 0$ and a k-dimensional subspace $E \subset \mathbb{R}^N$ such that

$$(1-\varepsilon) r D_E \subset K \cap E \subset (1+\varepsilon) r D_E \ ?$$

In fact, it has been conjectured (see [55]) that the answer to Question 1.2 is affirmative, with $c(k) = \frac{k-1}{2}$. This was proven by Bourgain and Lindenstrauss [20], for $k \geq 4$ and up to a factor of $\log \frac{1}{\varepsilon}$, but only when the convex set K is assumed to have certain symmetries. Question 1.2 has not even been resolved for small values of k; in particular $k = 3$. An exception is the case $k = 2$, where a proof due to Gromov appears in [55].

Question 1.1, Question 1.2 and problems of the same spirit are sensitive to the fine geometry of the convex body K. In this sense, these questions are more related to classical convexity theory. Moreover, the answers to the above two questions are both negative, if we relax the exact convexity requirement to quasi convexity, or even to isomorphic convexity (i.e., if we only assume that the convex hull of K is contained in $2K$).

We expect that in order to better understand the almost-isometric nature of high-dimensional convex bodies, new techniques should be employed, beyond the traditional concentration of measure phenomenon. Those techniques should take into account the precise convexity of the bodies, unlike in the isomorphic theory. Next, we demonstrate the transition from isomorphic to almost-isometric behavior in a specific test problem, that of geometric symmetrization.

2. Symmetrization of convex bodies

Let $K \subset \mathbb{R}^n$ be a convex body. For any hyperplane H that passes through the origin in $\mathbb{R}^n$ we will consider two types of symmetrization procedures. Our first symmetrization technique was described by Steiner ([66], see also [13]) in his proof of the isoperimetric inequality in two and three dimensions. Let $h \in S^{n-1}$ be a unit vector such that $H = h^{\perp}$. The Steiner symmetral of K with respect to the hyperplane H is the unique set $\sigma_H(K)$ for which the following two conditions hold:

1. For any $y \in H$, the set $\sigma_H(K) \cap [y + \mathbb{R}h]$ is a translation of the (possibly empty) segment $K \cap [y + \mathbb{R}h]$.

2. For any $y \in H$, the segment $\sigma_H(K) \cap [y + \mathbb{R}h]$, whenever non-empty, is centered at H.

Here $y + \mathbb{R}h$ stands for the line through y that is orthogonal to H. The set $\sigma_H(K)$ is symmetric with respect to the hyperplane H, hence the term "symmetrization". In addition, $\sigma_H(K)$ is convex and has the same volume as that of K. We will examine processes of symmetrization, where one begins with a convex body $K \subset \mathbb{R}^n$, and consecutively applies Steiner symmetrizations with respect to varying hyperplanes. It is a classical fact (see [25]) that given an arbitrary convex body $K \subset \mathbb{R}^n$, one may select appropriate hyperplanes $H_1, H_2, \ldots$ in $\mathbb{R}^n$ so that the sequence of bodies

$$\sigma_{H_m} \ldots \left(\sigma_{H_2}\left(\sigma_{H_1}(K)\right)\right) \quad \text{for } m = 1, 2, \ldots$$

converges in the Hausdorff metric to a Euclidean ball. This Euclidean ball will clearly have the same volume as that of the body we started with. Moreover, suppose we symmetrize a given convex body $K \subset \mathbb{R}^n$ with respect to randomly chosen hyperplanes, that are selected independently and uniformly over the grassmannian. Then convergence to a Euclidean ball occurs with probability one [49].

The second symmetrization procedure we consider is Minkowski symmetrization (also known as Blaschke symmetrization [9]). As before, $K \subset \mathbb{R}^n$ is a convex body and $H \subset \mathbb{R}^n$ is a hyperplane through the origin. For $x \in \mathbb{R}^n$, let $\pi_H(x)$ stand for the reflection of x with respect to H. The Minkowski symmetral of K with respect to H is the set

$$\tau_H(K) = \frac{K + \pi_H(K)}{2},$$

where $\frac{K+\pi_H(K)}{2} = \left\{\frac{x+\pi_H(y)}{2};\, x, y \in K\right\}$ is half of the Minkowski sum of K and $\pi_H(K)$. The set $\tau_H(K)$ is convex, yet its volume is usually different from that of K. Minkowski symmetrization preserves a different characteristic of the body, namely the mean width. The mean width of K is the quantity

$$w(K) = 2 \int_{S^{n-1}} \left[\sup_{x \in K} \langle x, \theta \rangle \right] d\mu(\theta),$$

where $S^{n-1} = \{x \in \mathbb{R}^n;\, |x| = 1\}$ is the unit sphere in $\mathbb{R}^n$ and μ is the unique rotationally-invariant probability measure on S^{n-1}. Thus, $w(\tau_H(K)) = w(K)$. A simple relation between Steiner and Minkowski symmetrization is that

$$\sigma_H(K) \subset \tau_H(K). \tag{2}$$

As in the case of Steiner symmetrizations, by applying an appropriate series of consecutive Minkowski symmetrizations to a given convex body $K \subset \mathbb{R}^n$, we obtain a sequence of convex bodies that converges towards a Euclidean ball. This Euclidean ball has the same mean width as the original body K.

Many geometric inequalities in which the Euclidean ball is the extremal case, may be proven using symmetrization techniques. Once we know that a certain geometric

quantity is, say, decreasing under symmetrization, we deduce that this quantity is minimized for the Euclidean ball, among all convex bodies of a given volume or mean width. For instance, from (2) we conclude that the ratio $w(K)/\operatorname{Vol}_n(K)^{\frac{1}{n}}$ is minimal for the Euclidean ball, among all convex bodies in $\mathbb{R}^n$. A sample of geometric inequalities proven via symmetrization includes the Brunn–Minkowski inequality (see, e.g., [13]), Santaló's inequality [50], Sylvester's problem [10], best approximation by polytopes [48] and a rearrangement inequality for integrals [23].

For a convex body $K \subset \mathbb{R}^n$ and $\varepsilon > 0$, we define $S(K,\varepsilon)$ (or $M(K,\varepsilon)$) to be the minimal number ℓ for which there exist ℓ Steiner symmetrizations (or Minkowski symmetrizations) that transform K into $\tilde{K}$ such that

$$e^{-\varepsilon} r D \subset \tilde{K} \subset e^{\varepsilon} r D,$$

where $D = \{x \in \mathbb{R}^n; |x| \leq 1\}$ is the unit Euclidean ball and $r = \left(\frac{\operatorname{Vol}_n(K)}{\operatorname{Vol}_n(D)}\right)^{\frac{1}{n}}$ (or $r = \frac{w(K)}{2}$). An interpretation I learned from V. Milman (e.g., [57]), is that the functions $S(K,\varepsilon)$, $M(K,\varepsilon)$ measure the complexity of the body K in the following sense. We view the Euclidean ball as the simplest of all convex bodies. If few symmetrizations are sufficient in order to transform K to become only ε-far from a Euclidean ball, then we think of K as being geometrically "simple". Convex bodies that require a large number of symmetrizations to attain this goal are viewed as more "complex". Define

$$S(n,\varepsilon) = \sup_{K \subset \mathbb{R}^n} S(K,\varepsilon), \quad M(n,\varepsilon) = \sup_{K \subset \mathbb{R}^n} M(K,\varepsilon), \tag{3}$$

where the suprema run over all convex bodies in $\mathbb{R}^n$. Consider first the isomorphic problem, where we try to symmetrize a convex body to make it close to a Euclidean ball in the isomorphic sense, as in (1). That is, we take ε in (3) to be of the order of magnitude of 1.

Theorem 2.1 ([37], [44]). *There exists a universal constant $c > 0$ such that for any dimension $n \geq 1$,*

1. *$S(n,c) \leq 3n$, and*
2. *$M(n,c) \leq 5n$.*

In addition, the slightly better inequality $M\left(n, c\frac{\log\log(n+2)}{\sqrt{\log(n+1)}}\right) \leq 5n$ holds.

Previous estimates in the literature are $M(n,c) < c' n \log n$ [21] and $S(n,c) < \tilde{c} n \log n$ [22]. See also [33] and [67]. Here, and throughout this note, the letters c, C, c', $\tilde{c}$ etc. denote positive universal constants. These constants need not be the same from one occurrence to the next. According to Theorem 2.1, all convex sets in $\mathbb{R}^n$ are geometrically "simple" in the above sense, at least in the isomorphic scale. Indeed, the number of symmetrizations needed to transform an arbitrary convex body into an isomorphic Euclidean ball, the simplest body, is only linear in the dimension n.

The constants "3" and "5" in Theorem 2.1 are probably not optimal, and the best constants are yet to be found. Yet, the exact constant is essentially known for a variant of our problem: Suppose we apply consecutive Steiner symmetrizations to a given convex body, and we are already satisfied when we arrive at an isomorphic ellipsoid, rather than a Euclidean ball. It is not very difficult to see (e.g. [44]) that for some convex bodies, at least $(1 - o(1))n$ Steiner symmetrizations are required in order to arrive at an isomorphic ellipsoid. The following theorem expresses the fact that roughly n symmetrizations are also sufficient, for all n-dimensional convex sets.

Theorem 2.2 ([44]). *For any $\delta > 0$, there exists a number $c(\delta) > 0$ for which the following holds: For any dimension $n \geq 1$ and a convex body $K \subset \mathbb{R}^n$, there exist an ellipsoid $\mathcal{E} \subset \mathbb{R}^n$ and $\lfloor (1+\delta)n \rfloor$ Steiner symmetrizations that transform K into a convex body $\tilde{K}$ such that*

$$\frac{1}{c(\delta)} \mathcal{E} \subset \tilde{K} \subset c(\delta) \mathcal{E}.$$

Moreover, $c(\delta) < c' \frac{1}{\delta} \log \frac{1}{\delta}$, where c' is a universal constant.

The proofs of Theorem 2.1 and Theorem 2.2 utilize some of the cornerstones of the asymptotic theory of convex geometry, such as concentration of measure inequalities, Kashin's splitting ([35], and the precise estimates in [30]) and Milman's quotient of subspace theorem mentioned above. Let us discuss some details from the proof of Theorem 2.1. We will focus our attention on the case of Minkowski symmetrizations, which is easier to analyze.

Given a convex body $K \subset \mathbb{R}^n$, our task is to design a sequence of symmetrizations that transform K into an approximate Euclidean ball. A plausible solution is choosing the symmetrizations randomly, that is, the hyperplanes are selected independently and uniformly. This approach was manifested in [21], and leads to the bound $M(n, c) < c' n \log n$. In fact, the effect of random Minkowski symmetrizations may be described even more precisely: For any convex body $K \subset \mathbb{R}^n$, the minimal number of *random* Minkowski symmetrizations needed in order to transform K, with reasonable probability, into an isomorphic Euclidean ball, has the order of magnitude of

$$n \log \frac{\mathrm{diam}(K)}{w(K)}. \tag{4}$$

Here $\mathrm{diam}(K)$ is the diameter of K (See [36] for exact formulation of this statement, based on [21]. See also [36] for a related phase-transition of the diameter in the process). We would like to emphasize that (4) is not merely a bound; it is actually an asymptotic formula for the minimal number of random symmetrizations required, valid for each convex body in $\mathbb{R}^n$. Just two simple geometric parameters, the diameter and the mean width, suffice to completely characterize the performance of a complicated process such as random Minkowski symmetrizations. This is a typical situation in asymptotic convex geometry (compare with [56] and [59]). The ratio $\frac{\mathrm{diam}(K)}{w(K)}$ is never larger than $c\sqrt{n}$, when $K \subset \mathbb{R}^n$. There are convex bodies in $\mathbb{R}^n$ for which

$\frac{\mathrm{diam}(K)}{w(K)} > c'\sqrt{n}$; a segment in $\mathbb{R}^n$ is an example of such a body. We thus conclude from (4) that for some convex bodies $K \subset \mathbb{R}^n$, at least $cn \log n$ random Minkowski symmetrizations are necessary in order to transform, with reasonable probability, the body K into an isomorphic Euclidean ball.

Consequently, the proof of the estimate $M(n, c) \leq 5n$ must involve a different symmetrization process: It is not efficient to simply take random, independent, Minkowski symmetrizations. This is in contrast to some other results in the theory, where the random choice is essentially the best choice (see, e.g. [59]). The approach taken in [37] is to perform several iterations, each consisting of n symmetrizations, that are carried out with respect to n mutually orthogonal hyperplanes in $\mathbb{R}^n$. There is more than one way of selecting these n mutually orthogonal hyperplanes. For instance, one may choose them randomly; that is, the iterations are independent, and the choice of the hyperplanes corresponds to the uniform probability measure on the orthogonal group (this leads to a proof that $M(n, c) \leq 6n$). As in [21], the proof of Theorem 2.1 still involves randomness, but of a different type.

We have explained why high-dimensional convex bodies are "simple objects", in some sense, in the isomorphic scale. One might be tempted to believe that the true complexity of high-dimensional convex sets resides in the almost-isometric scale. Perhaps the simplicity of convex bodies, as manifested in Theorem 2.1 and in the prominent theorems of asymptotic convex geometry (e.g., Theorem 1.1 and Theorem 1.2 above), is relevant only in the isomorphic scale? For the case of symmetrization, a negative answer is provided by the next theorem.

Theorem 2.3 ([39]). *There exists a universal constant $c > 0$ such that for any dimension $n \geq 1$ and $0 < \varepsilon < \frac{1}{2}$,*

1. $M(n, \varepsilon) \leq cn \log \frac{1}{\varepsilon}$, *and*
2. $S(n, \varepsilon) \leq cn^4 \log^2 \frac{1}{\varepsilon}$.

The proof of Theorem 2.3 involves harmonic analysis on the sphere S^{n-1}. The dependence on n and the dependence on ε in the bound for $M(n, \varepsilon)$ are each optimal, up to the exact value of the constant c. The exponents "4" and "2" in the bound for $S(n, \varepsilon)$ are probably not optimal. Yet, the dependence on ε in Theorem 2.3 is surprisingly good. Very few results with a logarithmic dependence on the distance ε are known in high-dimensional convex geometry. Another example is described in [7] (see S. Szarek's contribution in these proceedings for an explanation). It would be interesting to also find such good dependencies in other problems in the theory.

3. Volume distribution in convex bodies

The next family of problems we consider is related to the distribution of mass in high-dimensional convex bodies. For a convex body $K \subset \mathbb{R}^n$, let $\mathcal{E} \subset \mathbb{R}^n$ be the

Legendre ellipsoid of inertia of K; that is, $\mathcal{E}$ is the unique ellipsoid that has the same barycenter as K and also

$$\int_K \langle x, \theta \rangle^2 dx = \int_{\mathcal{E}} \langle x, \theta \rangle^2 dx$$

for any $\theta \in \mathbb{R}^n$. A convex body $K \subset \mathbb{R}^n$ of volume one is called isotropic if its barycenter lies at the origin and its Legendre ellipsoid is a Euclidean ball. In that case,

$$\int_K \langle x, \theta \rangle^2 dx = L_K^2$$

independently of θ in the unit sphere S^{n-1}. The quantity L_K is the isotropic constant of the convex body K. For any convex body $K \subset \mathbb{R}^n$ there exists a unique, up to orthogonal transformations, isotropic body $\tilde{K}$ which is an affine image of K (see, e.g., [51]). The isotropic constant of a general convex body K is defined as $L_K := L_{\tilde{K}}$, where $\tilde{K}$ is an isotropic affine image of K.

The isotropic constant of K encompasses many of the volumetric properties of the convex body K. See [51] for a list. For instance, if K is isotropic, then for any hyperplane H through the origin,

$$\frac{c_1}{L_K} \leq \mathrm{Vol}_{n-1}(K \cap H) \leq \frac{c_2}{L_K} \tag{5}$$

where $c_1, c_2 > 0$ are universal constants (see [34], or the survey paper [51]). Note that the relation (5) is a non-trivial rigidity property of convex bodies, in the almost-isometric scale. It is well-known (e.g. [51]) that for any dimension n and a convex body $K \subset \mathbb{R}^n$, we have $L_K > c$ for some universal constant $c > 0$. Denote,

$$L_n = \sup_{K \subset \mathbb{R}^n} L_K$$

where the supremum runs over all convex bodies in $\mathbb{R}^n$. Question 1.1 is equivalent (see [51]) to the following question: Is it true that $\sup_{n \geq 1} L_n < \infty$? The best estimate for the isotropic constant known to date is

$$L_n < c n^{\frac{1}{4}} \tag{6}$$

for a universal constant $c > 0$. The estimate (6), proven in [41], is a slight improvement on a previous bound $L_n < c n^{1/4} \log n$, due to Bourgain (See [16], [17], [28]. See [60] or the last remark in [38] for the non-symmetric case of Bourgain's bound). Aside from the general bound (6), an affirmative answer to Question 1.1 was obtained for large classes of convex bodies, including unconditional convex sets, zonoids, duals to zonoids, convex bodies with a bounded outer volume ratio and unit balls of Schatten norms (see, e.g., references in [41]). A reduction of the slicing problem, from general convex bodies to the simpler class of finite-volume-ratio bodies, appears in [18], [19] (see [18], [19] for precise definitions and statements).

A possible relaxation of Question 1.1 is its isomorphic version. Rather than trying to bound the isotropic constant of a given convex body $K \subset \mathbb{R}^n$, the isomorphic version asks whether there exists another convex body K', isomorphic to K in the sense of (1), for which the isotropic constant is bounded. A positive answer is provided in the following theorem.

Theorem 3.1 ([41]). *Let $K \subset \mathbb{R}^n$ be a convex body, and let $0 < \varepsilon < 1$. Then there exists another convex body $K' \subset \mathbb{R}^n$ such that*

1. $L_{K'} < \frac{c}{\sqrt{\varepsilon}}$, *and*
2. *for some $x_0 \in \mathbb{R}^n$,*

$$(1-\varepsilon)K' \subset K + x_0 \subset (1+\varepsilon)K'.$$

Here, $c > 0$ is a universal constant.

Theorem 3.1 reduces the slicing problem to a question regarding the stability of the isotropic constant under isomorphic change of the body. Theorem 3.1, together with Ball's observation (see [4] or [51, page 78]), provides another derivation of the existence of a Milman ellipsoid with a universal constant, for any convex body $K \subset \mathbb{R}^n$. A Milman ellipsoid for K with constant c is an ellipsoid $\mathcal{E} \subset \mathbb{R}^n$ with $\mathrm{Vol}_n(K) = \mathrm{Vol}_n(\mathcal{E})$ such that K may be covered by e^{cn} translations of $\mathcal{E}$ (see, e.g., [56] for a detailed discussion).

Given a convex body $K \subset \mathbb{R}^n$, there are several ellipsoids or Euclidean structures associated with K, such as Milman ellipsoids, the maximal volume ellipsoid, the Legendre inertia ellipsoid, the minimal surface area ellipsoid, etc. It is customary to call these Euclidean structures various "positions" of K. The relations between different positions of a convex body are not clear in general. See [43] for a certain non-trivial relation, applicable only to 2-convex bodies. As is proven in [19], Question 1.1 is equivalent to the following question: Is it true that for any convex body $K \subset \mathbb{R}^n$, the Legendre ellipsoid of K is also a Milman ellipsoid for K, with a universal constant?

A very interesting development stems from the recent Paouris theorem. Suppose that $K \subset \mathbb{R}^n$ is an isotropic convex body. Let X be a random vector, that distributes uniformly over K. Then $\mathbb{E}X = 0$. What can be said about the distribution of $|X|$? Clearly $\sqrt{\mathbb{E}|X|^2} = \sqrt{n}L_K$. Moreover, a direct consequence of the Brunn–Minkowski inequality is that $\mathrm{Prob}(|X| > \sqrt{n}L_K t)$ decays exponentially in t, i.e., at least as fast as e^{-ct} for a universal constant $c > 0$ (see [1] for a subgaussian decay). A surprisingly strong improvement is contained in the following theorem [63], [64].

Theorem 3.2 (Paouris theorem). *Let $K \subset \mathbb{R}^n$ be an isotropic convex body. Then, for any $t \geq 1$,*

$$\mathrm{Vol}_n\left(\{x \in K; |x| \geq ct\sqrt{n}L_K\}\right) \leq \exp(-t\sqrt{n}) \tag{7}$$

where $c > 0$ is a universal constant.

The proof of Theorem 3.2 involves, among other ingredients, a clever use of Dvoretzky's theorem. In the case where the convex body K is also assumed to be unconditional, the conclusion of Theorem 3.2 was proven in [11], [12], and for K being the normalized ℓ_1^n-ball, the result was proven in [65]. The inequality (7) is actually tight for the normalized ℓ_1^n-ball, up to the value of the constant c.

According to Theorem 3.2, all the mass of an isotropic convex body, except for a mere $e^{-\sqrt{n}}$-fraction, lies inside a ball of radius $c\sqrt{n}L_K$ around the origin. A conjecture put forward by Anttila, Ball and Perissinaki [2] suggests that there exists a sequence $\varepsilon_n \to 0$ with the following property: Whenever $K \subset \mathbb{R}^n$ is an isotropic convex body, then for some $\rho > 0$,

$$\mathrm{Vol}_n\left(\{x \in K;\, (1-\varepsilon_n)\rho \le |x| \le (1+\varepsilon_n)\rho\}\right) \ge 1 - \varepsilon_n. \tag{8}$$

This "thin shell" conjecture (8) was verified in [2] for unit balls of l_p^n-spaces, and for a large family of uniformly convex bodies. A positive answer to this conjecture would imply, in particular, that all high-dimensional isotropic convex bodies have many near-gaussian one-dimensional marginal distributions. See [2] for the exact formulation and proof of this implication, and see [24] for a discussion pertaining to the question of existence of near-gaussian marginals, for all high-dimensional convex bodies.

Our next topic is related to large deviation estimates for marginal distributions of general convex sets. Suppose $K \subset \mathbb{R}^n$ is a convex body of volume one, and let $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ be a linear functional. Denote $\|\varphi\|_{L^1(K)} = \int_K |\varphi(x)|\,dx$. A well-known consequence of the Brunn–Minkowski inequality, observed by Borell [14], is that for all $t \ge 1$,

$$\mathrm{Vol}_n\left(\{x \in K;\, |\varphi(x)| \ge t\|\varphi\|_{L^1(K)}\}\right) \le \exp(-ct) \tag{9}$$

where $c > 0$ is a universal constant. Thus, a uniform sub-exponential estimate holds for the distribution of an arbitrary linear functional on an arbitrary convex set. A typical case in which (9) is sharp, is that of a cone over an $(n-1)$-dimensional base; the distribution of a linear functional that vanishes on the base of the cone, is very close to being an exact exponential.

When $K \subset \mathbb{R}^n$ is an ellipsoid of volume one, the sub-exponential bound (9) may be substantially improved. It is easy to see that in this case, any linear functional $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ satisfies the sub-gaussian estimate

$$\mathrm{Vol}_n\left(\{x \in K;\, |\varphi(x)| \ge t\|\varphi\|_{L^1(K)}\}\right) \le \exp(-ct^2) \quad \text{for all } t \ge 1, \tag{10}$$

where $c > 0$ is a universal constant. Inequality (10) is rather sharp, since the distribution of a linear functional on an ellipsoid is very close to being gaussian. A question that is often attributed to Milman [6], [61], [62], asks whether for any convex body $K \subset \mathbb{R}^n$ of volume one, there exists a non-zero linear functional $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, for which a sub-gaussian estimate holds as in (10). A positive answer to this question would have the interpretation that for any convex body $K \subset \mathbb{R}^n$, there exists a direction in which, in a sense, K does not look like an apex or a cone, but rather exhibits quite regular behavior, like that of an ellipsoid or a Euclidean ball.

An affirmative answer to Milman's question was obtained for unconditional convex bodies [11], for zonoids [61] and for some other classes of convex sets [61], [62]. A recent, general principle provides an affirmative answer to Milman's question, up to a logarithmic factor:

Theorem 3.3 ([42]). *Let $K \subset \mathbb{R}^n$ be a convex body of volume one. Then there exists a non-zero linear functional $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, such that for any $t \geq 1$,*

$$\mathrm{Vol}_n\left(\{x \in K;\, |\varphi(x)| \geq t\|\varphi\|_{L^1(K)}\}\right) \leq \exp\left(-c\,\frac{t^2}{\log^5(t+1)}\right),$$

where $c > 0$ is a universal constant.

The proofs of Theorem 3.1 and Theorem 3.3 make use of several properties of the logarithmic Laplace transform of log-concave functions. We would like to conclude this section with the "random cotype-2" result of Gluskin and Milman. Suppose $K \subset \mathbb{R}^n$ is a centrally-symmetric convex body, and $X_1, \ldots, X_n$ are independent, random vectors, distributed uniformly in K. In [32] it is proven that with probability larger than $1 - e^{-cn}$,

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\left\|\sum_{i=1}^n \varepsilon_i\lambda_i X_i\right\|_K > c\sqrt{\sum_{i=1}^n \lambda_i^2} \quad \text{for all } (\lambda_1,\ldots,\lambda_n) \in \mathbb{R}^n, \tag{11}$$

where $\|\cdot\|_K$ is the norm whose unit ball is K and $c > 0$ is a universal constant. Consequently, any finite-dimensional norm satisfies a cotype-2 condition as in (11), with high probability, when the vectors $X_1, \ldots, X_n$ are random vectors that are selected independently and uniformly in the unit ball of that norm. See, e.g., [58, Section 9], for definitions and basic properties of type and cotype of normed spaces.

4. Beyond Brunn–Minkowski and Santaló inequalities

Some of the recent developments regarding volume distribution in high-dimensional convex sets are connected with a better understanding of log-concave functions, which are functions $f\colon \mathbb{R}^n \to [0,\infty)$ whose logarithm is concave. The relation between the slicing problem and log-concave functions goes back at least to [5]. Recall that the Legendre transform of a function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\mathcal{L}\varphi(x) = \sup_{y\in\mathbb{R}^n}\left[\langle x, y\rangle - \varphi(y)\right].$$

The following result follows from the Santaló and Bourgain–Milman inequalities [3], [45]: For any measurable function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, there exists $x_0 \in \mathbb{R}$ such that $\tilde{\varphi}(x) = \varphi(x - x_0)$ satisfies

$$c \leq \left(\int_{\mathbb{R}^n} e^{-\tilde{\varphi}} \int_{\mathbb{R}^n} e^{-\mathcal{L}\tilde{\varphi}}\right)^{\frac{1}{n}} \leq 2\pi, \tag{12}$$

where $c > 0$ is a universal constant (and we agree, for the purpose of (12), that $c \leq \infty \cdot 0 \leq 2\pi$). Equality on the right hand side of (12) holds if and only if φ is a.e. a positive definite quadratic form (see [3]). For the case where φ is an even function, we may select $x_0 = 0$ in (12). In that case, the right hand side of (12) was proven by K. Ball (see [4] and also [29], for related inequalities). When φ is an even function, the following generalization holds (see [40]):

$$\int_{\mathbb{R}^n} e^{-\varphi}\, d\mu \int_{\mathbb{R}^n} e^{-\mathcal{L}\varphi}\, d\mu \leq \left(\int_{\mathbb{R}^n} e^{-\frac{|x|^2}{2}}\, d\mu \right)^2 \tag{13}$$

where μ is any log-concave measure on $\mathbb{R}^n$ (for example, a measure on $\mathbb{R}^n$ whose density is a log-concave function). This is closely related to an interesting theorem of Cordero-Erausquin [26]: Suppose $K, T \subset \mathbb{R}^{2n}$ are convex bodies. We endow $\mathbb{R}^{2n}$ with a complex structure, and assume that K, T are unit balls of complex Banach norms, and $T = \overline{T}$ where $\overline{T}$ is the conjugate of T. Then

$$\mathrm{Vol}_{2n}(K \cap T)\, \mathrm{Vol}_{2n}(K^\circ \cap T) \leq \mathrm{Vol}_{2n}(D \cap T)^2, \tag{14}$$

where $K^\circ = \{x \in \mathbb{R}^{2n};\ \text{for all } y \in K,\ \langle x, y\rangle \leq 1\}$ is the dual body. The proof of (14) uses complex interpolation and a recent complex version of the Prékopa–Leindler inequality due to Berndtsson [8]. It is not clear at the moment whether (14) generalizes to arbitrary centrally-symmetric convex sets. Inequality (13) may be viewed as a functional version of this suspected generalization. See [40] for related inequalities.

Inequality (14) suggests that, perhaps, convex bodies obey some additional geometric inequalities, beyond the classical Santaló and Brunn–Minkowski inequalities. Further evidence for this stems from the result of Cordero-Erausquin, Fradelizi and Maurey in [27]. Solving a conjecture from [46], they show that for any centrally-symmetric convex body $K \subset \mathbb{R}^n$ and $s, t > 0$,

$$\gamma_n(\sqrt{st}K) \geq \sqrt{\gamma_n(sK)\gamma_n(tK)}, \tag{15}$$

where γ_n is the standard gaussian measure in $\mathbb{R}^n$, whose density is given by $d\gamma_n = (2\pi)^{-n/2} \exp(-|x|^2/2)dx$. The Brunn–Minkowski type arguments only yield (15) with $\sqrt{st}$ replaced by $\frac{s+t}{2}$ (see also F. Barthe's article in this volume). In the case where n is even, and K is the unit ball of a complex Banach norm, it is possible to replace the gaussian measure in (15) with any log-concave measure that respects the complex structure in a natural way (this follows from [26, Theorem 3.2]). It would be desirable to understand whether (14) and (15) actually hold in the context of arbitrary centrally-symmetric convex sets and arbitrary even log-concave measures, without an underlying complex structure.

Note added in proof. We would like to report on two very recent developments: First, Giannopoulos, Pajor and Paouris have simplified and slightly improved the proof of

Theorem 3.3, see http://arxiv.org/abs/math.FA/0604299. Second, the "thin shell" conjecture (8) has been proved by the author for all isotropic, convex sets. Consequently, typical one-dimensional marginal distributions of high-dimensional, isotropic, convex sets are approximately gaussian. Similar principles also hold for multi-dimensional marginal distributions. See http://arxiv.org/abs/math.MG/0605014.

References

[1] Alesker, S., ψ_2-estimate for the Euclidean norm on a convex body in isotropic position. In *Geometric aspects of functional analysis*, Israel Seminar (1992–1994), Oper. Theory Adv. Appl. 77, Birkhäuser, Basel 1995, 1–4.

[2] Anttila, M., Ball, K., Perissinaki, I., The central limit problem for convex bodies. *Trans. Amer. Math. Soc.* **355** (12) (2003), 4723–4735.

[3] Artstein-Avidan, S., Klartag, B., Milman, V., The Santaló point of a function, and a functional form of Santaló inequality. *Mathematika*, to appear.

[4] Ball, K., Isometric problems in ℓ_p and sections of convex sets. Ph.D. Dissertation, Trinity College, Cambridge, 1986.

[5] Ball, K., Logarithmically concave functions and sections of convex sets in $\mathbb{R}^n$. *Studia Math.* **88** (1) (1988), 69–84.

[6] Barthe, F., Guédon, O., Mendelson, S., Naor, A., A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.* **33** (2005), 480–513.

[7] Ben-Tal, A., A. Nemirovski, On polyhedral approximations of the second-order cone *Math. Oper. Res.* **26** (2) (2001), 193–205.

[8] Berndtsson, B., Prékopa's theorem and Kiselman's minimum principle for plurisubharmonic functions. *Math. Ann.* **312** (1998), 785–792.

[9] Blaschke, W., *Kreis und Kugel.* Veit, Leipzig 1916, 103–104.

[10] Blaschke, W., Über affine Geometrie XI: Lösung des "Vierpunktproblems" von Sylvester aus der Theorie der geometrischen Wahrscheinlichkeiten. *Leipziger Berichte* **69** (1917), 436–453.

[11] Bobkov, S. G., Nazarov, F. L., On convex bodies and log-concave probability measures with unconditional basis. In *Geometric aspects of functional analysis*, Israel Seminar (2001–2002), Lecture Notes in Math. 1807, Springer-Verlag, Berlin 2003, 53–69.

[12] Bobkov, S. G., Nazarov, F. L., Large deviations of typical linear functionals on a convex body with unconditional basis. In *Stochastic inequalities and applications*, Progr. Probab. 56, Birkhäuser, Basel 2003, 3–13.

[13] Bonnesen, T., Fenchel, W., *Theory of convex bodies*. English transl. of the German original (Erg. Math. Grenzgeb. 3, Springer-Verlag, Berlin 1934), BCS Associates, Moscow, ID, 1987.

[14] Borell, C., The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30** (2) (1975), 207–216.

[15] Bourgain, J., Geometry of Banach spaces and harmonic analysis. In *Proceedings of the International Congress of Mathematicians* (Berkeley, Calif., 1986), Vol. 2, Amer. Math. Soc., Providence, RI, 1987, 871–878.

[16] Bourgain, J., On the distribution of polynomials on high-dimensional convex sets. In *Geometric aspects of functional analysis*, Israel Seminar (1989–90), Lecture Notes in Math. 1469, Springer-Verlag, Berlin 1991, 127–137.

[17] Bourgain, J., On the isotropy-constant problem for "PSI-2"-bodies. In *Geometric aspects of functional analysis*, Israel Seminar (2001–2002), Lecture Notes in Math. 1807, Springer-Verlag, Berlin 2003, 114–121.

[18] Bourgain, J., Klartag, B., Milman, V., A reduction of the slicing problem to finite volume ratio bodies. *C. R. Math. Acad. Sci. Paris* **336** (4) (2003), 331–334.

[19] Bourgain, J., Klartag, B., Milman, V., Symmetrization and isotropic constants of convex bodies. In *Geometric aspects of functional analysis*, Israel Seminar (2002–2003), Lecture Notes in Math. 1850, Springer-Verlag, Berlin 2004, 101–115.

[20] Bourgain, J., Lindenstrauss, J., Almost Euclidean sections in spaces with a symmetric basis. In *Geometric aspects of functional analysis*, Israel Seminar (1987–88), Lecture Notes in Math. 1376, Springer-Verlag, Berlin 1989, 278–288.

[21] Bourgain, J., Lindenstrauss, J., Milman, V., Minkowski sums and symmetrizations. In *Geometric aspects of functional analysis*, Israel Seminar (1986–87), Lecture Notes in Math. 1317, Springer-Verlag, Berlin 1988, 44–66.

[22] Bourgain, J., Lindenstrauss, J., Milman, V., Estimates related to Steiner symmetrizations. In *Geometric aspects of functional analysis*, Israel Seminar (1987–88), Lecture Notes in Math. 1376, Springer-Verlag, Berlin 1989, 264–273.

[23] Brascamp, H. J., Lieb, E. H., Luttinger, J. M., A general rearrangement inequality for multiple integrals. *J. Funct. Anal.* **17** (1974), 227–237.

[24] Brehm, U., Voigt, J., Asymptotics of cross sections for convex bodies. *Beiträge Algebra Geom.* **41** (2) (2000), 437–454.

[25] Carathéodory, C., Study, E., Zwei Beweise des Satzes, daß der Kreis unter allen Figuren gleichen Umfanges den größten Inhalt hat. *Math. Ann.* **68** (1910), 133–140.

[26] Cordero-Erausquin, D., Santaló's inequality on $\mathbb{C}^n$ by complex interpolation. *C. R. Math. Acad. Sci. Paris* **334** (9) (2002), 767–772.

[27] Cordero-Erausquin, D., Fradelizi, M., Maurey, B., The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems. *J. Funct. Anal.* **214** (2) (2004), 410–427.

[28] Dar, S., Remarks on Bourgain's problem on slicing of convex bodies. In *Geometric aspects of functional analysis,* Israel Seminar (1992–1994), Oper. Theory Adv. Appl. 77, Birkhäuser, Basel 1995, 61–66.

[29] Fradelizi, M., Meyer, M., Some functional forms of Blaschke-Santaló inequality. Preprint.

[30] Garnaev, A., Gluskin, E., The widths of a Euclidean ball. *Dokl. Akad. Nauk SSSR* **277** (5) (1984) 1048–1052; English transl. *Soviet Math. Dokl.* **30** (1) (1984), 200–204.

[31] Giannopoulos, A., Milman, V., Euclidean structure in finite dimensional normed spaces. In *Handbook of the geometry of Banach spaces*, Vol. I, North-Holland, Amsterdam 2001, 707–779.

[32] Gluskin, E., Milman, V., Geometric probability and random cotype 2. In *Geometric aspects of functional analysis*, Israel Seminar (2002–2003) Lecture Notes in Math. 1850, Springer-Verlag, Berlin 2004, 123–138.

[33] Hadwiger, H., Einfache Herleitung der isoperimetrischen Ungleichung für abgeschlossene Punktmengen. *Math. Ann.* **124** (1952), 158–160.

[34] Hensley, D., Slicing convex bodies—bounds for slice area in terms of the body's covariance. *Proc. Amer. Math. Soc.* **79** (4) (1980), 619–625.

[35] Kashin, B. S., Diameters of some finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk. SSSR. Ser. Mat.* **41** (2) (1977), 334–351; English transl. *Math. USSR Izv.* **11** (2) (1977), 317–333.

[36] Klartag, B., Remarks on Minkowski Symmetrizations. In *Geometric aspects of functional analysis*, Israel Seminar (1996–2000), Lecture Notes in Math., Vol. 1745, Springer, Berlin 2000, 109–118.

[37] Klartag, B., 5n Minkowski symmetrizations suffice to arrive at an approximate Euclidean ball. *Ann. of Math.* **156** (3) (2002), 947–960.

[38] Klartag, B., An isomorphic version of the slicing problem. *J. Funct. Anal.* **218** (2) (2005), 372–394.

[39] Klartag, B., Rate of convergence of geometric symmetrization. *Geom. Funct. Anal*. **14** (6) (2004), 1322–1338.

[40] Klartag, B., Marginals of geometric inequalities. To appear in *Geometric aspects of functional analysis*, Israel Seminar (2004–2005).

[41] Klartag, B., On convex perturbations with a bounded isotropic constant. *Geom. Funct. Anal.*, to appear.

[42] Klartag, B., Uniform almost sub-gaussian estimates for linear functionals on convex sets. Preprint.

[43] Klartag, B., Milman, E., On volume distribution in 2-convex bodies. In preparation.

[44] Klartag, B., Milman, V., Isomorphic Steiner symmetrizations. *Invent. Math.* **153** (3) (2003), 463–485.

[45] Klartag, B., Milman, V., Geometry of log-concave functions and measures. *Geom. Dedicata* **112** (2005), 169–182.

[46] Latała, R., On some inequalities for Gaussian measures. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. II, Higher Ed. Press, Beijing 2002, 813–822.

[47] Lindenstrauss, J., Milman, V., The local theory of normed spaces and its applications to convexity. In *Handbook of convex geometry*, Vol. B, North-Holland, Amsterdam 1993, 1149–1220.

[48] Macbeath, A. M., An extremal property of the hypersphere. *Proc. Cambridge Philos. Soc.* **47** (1951), 245–247.

[49] Mani-Levitska, P., Random Steiner symmetrizations. *Studia Sci. Math. Hung.* **21** (1986), 373–378.

[50] Meyer, M., Pajor, A., On the Blaschke-Santalo inequaqlity. *Arch. Math.* **55** (1990), 82–93.

[51] Milman, V., Pajor, A., Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space. In *Geometric aspects of functional analysis*, Israel Seminar (1987–88), Lecture Notes in Math. 1376, Springer-Verlag, Berlin 1989, 64–104.

[52] Milman, V., A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.* **5** (4) (1971), 28–37; English transl. *Funct. Anal. Appl.* **5** (1971), 288–295.

[53] Milman, V., Almost Euclidean qoutient spaces of subspaces of a finite dimensional normed space. *Proc. Amer. Math. Soc.* **94** (1985), 445–449.

[54] Milman, V., The concentration phenomenon and linear structure of finite-dimensional normed spaces. In *Proceedings of the International Congress of Mathematicians* (Berkeley, Calif., 1986), Vol. 2, Amer. Math. Soc., Providence, RI, 1987, 961–975.

[55] Milman, V., A few observations on the connections between local theory and some other fields. In *Geometric aspects of functional analysis*, Israel Seminar (1986–87), Lecture Notes in Math. 1317, Springer-Verlag, Berlin 1988, 283–289.

[56] Milman, V., Randomness and pattern in convex geometric analysis. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 665–677.

[57] Milman, V., Unified behavior of convex bodies in high dimensional spaces. Conference talk, Barcelona, June 2004. Slides available at http://www.crm.es/Conferences/0304/LearningTheory/Lectures/slides31-5-04milman.pdf.

[58] Milman, V., Schechtman, G., *Asymptotic theory of finite-dimensional normed spaces.* Lecture Notes in Math. 1200, Springer-Verlag, Berlin 1986.

[59] Milman, V., Schechtman, G., Global versus local asymptotic theories of finite-dimensional normed spaces. *Duke Math. J.* **90** (1) (1997), 73–93.

[60] Paouris, G., On the isotropic constant of non-symmetric convex bodies. In *Geometric aspects of functional analysis*, Israel Seminar (1996–2000), Lecture Notes in Math. 1745, Springer-Verlag, Berlin 2000, 239–243.

[61] Paouris, G., Ψ_2-estimates for linear functionals on zonoids. In *Geometric aspects of functional analysis,* Israel Seminar (2001–2002), Lecture Notes in Math. 1807, Springer-Verlag, Berlin 2003, 211–222.

[62] Paouris, G., On the Ψ_2 behavior of linear functionals on isotropic convex bodies. *Studia Math.* **168** (2005), 285-299.

[63] Paouris, G., Concentration of mass on isotropic convex bodies. *C. R. Math. Acad. Sci. Paris.* **342** (3) (2006), 179–182.

[64] Paouris, G., Concentration of mass on convex bodies. Preprint.

[65] Schechtman, G., Zinn, J., On the volume of the intersection of two L_p^n balls. *Proc. Amer. Math. Soc.* **110** (1) (1990), 217–224.

[66] Steiner, J., Einfacher Beweis der isoperimetrischen Hauptsätze. *J. Reine Angew. Math.* **18** (1838), 281–296; Gesammelte Werke 2, G. Reimer, Berlin 1882, 77–91; reprint by Chelsea, Bronx, NY, 1972.

[67] Tsolomitis, A., Quantitative Steiner/Schwarz-type symmetrization. *Geom. Dedicata* **60** (2) (1996), 187–206.

School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, U.S.A.
E-mail: klartag@ias.edu

Amenable actions and applications

Narutaka Ozawa

Abstract. We will give a brief account of (topological) amenable actions and exactness for countable discrete groups. The class of exact groups contains most of the familiar groups and yet is manageable enough to provide interesting applications in geometric topology, von Neumann algebras and ergodic theory.

Mathematics Subject Classification (2000). Primary 46L35; Secondary 20F65, 37A20, 43A07.

Keywords. Amenable actions, exact groups, group von Neumann algebras.

1. Introduction

The notion of amenable groups was introduced by J. von Neumann in 1929 in his investigation of the Banach–Tarski paradox. He observed that non-abelian free groups are not amenable and that this fact is the source of the Banach–Tarski paradox. Since then it has been shown that the amenability of a locally compact group is equivalent to many fundamental properties in harmonic analysis of the group: the Følner property, the fixed point property and the weak containment of the trivial representation in the regular representation, to name a few. For a discrete group, amenability of the group is also characterized by nuclearity of its group C^*-algebra, and by injectivity of its group von Neumann algebra. In this note, we are mainly interested in countable discrete groups. The class of amenable groups contains all solvable groups and is closed under subgroups, quotients, extensions and directed unions. As we mentioned before, a non-abelian free group, or any group which contains it, is not amenable. Amenable groups play a pivotal role in the theory of operator algebras. Many significant operator algebra-related problems on groups have been solved for amenable groups. We just cite two of them; the classification of group von Neumann algebras [14] and measure equivalences [16], [58] on the one hand, and the Baum–Connes conjecture [46] on the other hand. In recent years, there have been exciting breakthroughs in both subjects beyond the amenable cases. We refer to [68] and [84] for accounts of this progress. We will also treat the classification of group von Neumann algebras and measure equivalences in Section 4. Since many significant problems, if not all, are already solved for amenable groups, we would like to set out for the world of non-amenable groups. Still, as Gromov's principle goes, no statement about all groups is both non-trivial and true. So we want a *good* class of groups to play with. We consider a class

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

as good if it contains many of the familiar examples, is manageable enough so that it maintains non-trivial theorems, and can be characterized in various ways so that it is versatile. We believe that the class of exact groups, which will be introduced in the following section, stands these tests. The study of exactness originates in C^*-algebra theory [50], [51], [52] and was propagated to groups. The class of exact groups is fairly large and it contains all amenable groups, linear groups [39] and hyperbolic groups [2], to name a few. It is closed under subgroups, extensions, directed unions and amalgamated free products. (Since every free group is exact and there exists a non-exact group [37], a quotient of exact group needs not be exact unless the normal subgroup is amenable.) Moreover, there is a remarkable theorem that the injectivity part of the Baum–Connes conjecture holds for exact groups [45], [76], [83], [84]. Since this part of the Baum–Connes conjecture has a lot of applications in geometry and topology, including the strong Novikov conjecture, it is an interesting challenge to prove exactness of a given group. We will encounter some other applications in von Neumann algebra theory and ergodic theory in Section 4.

2. Amenable actions and exactness

We first review the definition of and basic facts on amenable actions. We refer to [65] for the theory of amenable groups and to [5], [11] for the theory of amenable actions. The notion of amenability for a group action was first introduced in the measure space setting in the seminal paper [85], which has had a great influence in both ergodic theory and von Neumann algebra theory. In this spirit the study of its topological counterpart was initiated in [3]. In this note, we restrict our attention to continuous actions of countable discrete groups on (not necessarily second countable) compact spaces. All topological spaces are assumed to be Hausdorff and all groups, written as $\Gamma, \Lambda, \dots$, are assumed to be countable and discrete. Let Γ be a group. A (topological) Γ-*space* is a topological space X together with a continuous action of Γ on it; $\Gamma \times X \ni (s, x) \mapsto s.x \in X$. For a group (or any countable set) Γ, we let

$$\mathrm{prob}(\Gamma) = \left\{\mu \in \ell_1(\Gamma) : \mu \geq 0,\ \textstyle\sum_{t\in\Gamma} \mu(t) = 1\right\} \subset \ell_1(\Gamma)$$

and equip $\mathrm{prob}(\Gamma)$ with the pointwise convergence topology. We note that this topology coincides with the norm topology. The space $\mathrm{prob}(\Gamma)$ is a Γ-space with the Γ-action given by the left translation: $(s.\mu)(t) = \mu(s^{-1}t)$.

Definition 2.1. We say that a compact Γ-space X is *amenable* (or Γ *acts amenably* on X) if there exists a sequence of continuous maps

$$\mu_n \colon X \ni x \mapsto \mu_n^x \in \mathrm{prob}(\Gamma)$$

such that for every $s \in \Gamma$ we have

$$\lim_{n\to\infty} \sup_{x\in X} \|s.\mu_n^x - \mu_n^{s.x}\| = 0.$$

When X is a point, the above definition degenerates to one of the equivalent definitions of amenability for the group Γ. Moreover, if Γ is amenable, then every Γ-space is amenable. Conversely, if there exists an amenable Γ-space which carries an invariant Radon probability measure, then Γ itself is amenable. If X is an amenable Γ-space, then X is amenable as a Λ-space for every subgroup Λ. It follows that all isotropy subgroups of an amenable Γ-space have to be amenable. We recall that the isotropy subgroup of x in a Γ-space X is $\{s \in \Gamma : s.x = x\}$. It is also easy to see that if there exists a Γ-equivariant continuous map from a Γ-space Y into another Γ-space X and if X is amenable, then so is Y. Finally, we only note that there are several equivalent characterizations of an amenable action which generalize those for an amenable group.

Many amenable actions naturally arise from the geometry of groups. The following are the most basic examples of amenable actions.

Example 2.2. Let $\mathbb{F}_r = \langle g_1, \ldots, g_r \rangle$ be the free group of rank $r < \infty$. Then its (Gromov) boundary $\partial\mathbb{F}_r$ is amenable. We note that the Cayley graph of $\mathbb{F}_r$ w.r.t. the standard set of generators is a simplicial tree and its boundary

$$\partial\mathbb{F}_r \subset \{g_1, g_1^{-1}, \ldots, g_r, g_r^{-1}\}^{\mathbb{N}}$$

is defined as the compact topological space of all infinite reduced words, equipped with the relative product topology (see Figure 1). Similarly, with an appropriate topology, $\mathbb{F}_r \cup \partial\mathbb{F}_r$ becomes a compactification of $\mathbb{F}_r$. The free group $\mathbb{F}_r$ acts continuously on $\partial\mathbb{F}_r$ by left multiplication (and rectifying possible redundancy). For $x \in \partial\mathbb{F}_r$ with its reduced form $x = a_1 a_2 \ldots$, we set $x_0 = e$ and $x_k = a_1 \ldots a_k$. For every $n \in \mathbb{N}$, we let

$$\mu_n \colon \partial\mathbb{F}_r \ni x \mapsto \mu_n^x = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{x_k} \in \mathrm{prob}(\mathbb{F}_r).$$

Thus μ_n^x is the normalized characteristic function of the first n segments of the path in the Cayley graph of $\mathbb{F}_r$, connecting e to x (see Figure 1). It is not hard to see that μ_n is a continuous map such that

$$\sup_{x \in \partial\mathbb{F}_r} \|s.\mu_n^x - \mu_n^{s.x}\| \leq \frac{2|s|}{n}$$

for every $s \in \mathbb{F}_r$, where $|s|$ is the word length of s. Indeed, $s.\mu_n^x$ is the normalized characteristic function of the first n segments of the path connecting s to $s.x$, which has a large intersection with the path connecting e to $s.x$ (see Figure 2).

There are generalizations of this construction to groups acting on more general buildings [72] and on hyperbolic spaces [2].

Example 2.3. Let Γ be a discrete subgroup of the special linear group $\mathrm{SL}(n, \mathbb{R})$ (e.g., $\Gamma = \mathrm{SL}(n, \mathbb{Z})$) and $P \subset \mathrm{SL}(n, \mathbb{R})$ be the closed subgroup of upper triangular matrices.

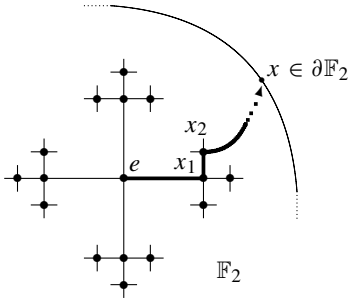


Figure 1. The Cayley graph of $\mathbb{F}_2$ and the boundary $\partial\mathbb{F}_2$.

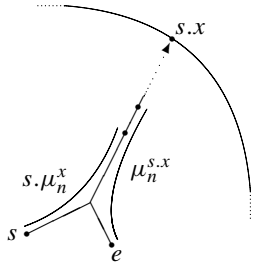


Figure 2. Amenability of $\partial\mathbb{F}_2$.

Then the left multiplication action of Γ on the Furstenberg boundary $\mathrm{SL}(n,\mathbb{R})/P$ is amenable. More generally, if G is a locally compact group with a closed amenable locally compact subgroup P (such that G/P compact), then every discrete subgroup Γ of G acts amenably on G/P.

A far-reaching generalization of this example is given in [39], where it is shown that any linear group admits an amenable action on some compact space. Thus many non-amenable groups admit amenable actions.

Definition 2.4. We say a group Γ is *exact* if there exists a compact Γ-space X which is amenable.

Exact groups are also said to be *boundary amenable*, *amenable at infinity* or to *have the property A*. By definition, all amenable groups are exact. Let X be a compact Γ-space. Then, by the universality of the Stone–Čech compactification $\beta\Gamma$, there exists a Γ-equivariant continuous map from $\beta\Gamma$ into X. It follows that Γ is exact iff $\beta\Gamma$ (or the boundary $\partial^\beta\Gamma = \beta\Gamma \setminus \Gamma$) is amenable. Moreover, whether Γ is exact or not, $\beta\Gamma$ is amenable as a Λ-space for every exact subgroup Λ of Γ since there exists a Λ-equivariant continuous map from $\beta\Gamma$ into $\beta\Lambda$. This observation implies that exactness is preserved under a directed union, i.e., a group Γ is exact iff all of its finitely generated subgroups are exact.

Amenability of the Stone–Čech compactification $\beta\Gamma$ leads to an intrinsic characterization of an exact group Γ. Before stating it, we introduce the notion of coarse metric spaces [36]. Let d be a left translation invariant metric on Γ which is proper in the sense that every subset of finite diameter is finite. Then $l(s) = d(s,e)$ is a length function on Γ, i.e., $l(s^{-1}) = l(s)$, $l(st) \leq l(s) + l(t)$ for every $s, t \in \Gamma$, and $l(s) = 0$ iff $s = e$. The length function l is proper in the sense that $l^{-1}([0, R])$ is finite for every $R > 0$. Conversely, every proper length function l gives rise to a proper left translation invariant metric d on Γ such that $d(s,t) = l(s^{-1}t)$. If $\mathcal{S}$ is a finite

generating subset of Γ, then the corresponding word metric is defined by

$$d_{\mathcal{S}}(s,t) = \min\{n : s^{-1}t = s_1 \dots s_n,\ s_i \in \mathcal{S} \cup \mathcal{S}^{-1}\}.$$

We note that even when Γ is not finitely generated, there exists a proper left translation invariant metric d on Γ (as we assume that Γ is countable). Two proper length functions l and l' are equivalent in the sense that $l(s_n) \to \infty$ iff $l'(s_n) \to \infty$. Thus we are lead to the notion of coarse equivalence, which is a very loose notion. Two metric spaces (X, d) and (X', d') are *coarsely isomorphic* if there exists a (not necessarily continuous) map $f\colon X \to X'$ such that $d(z, f(X)) < \infty$ for every $z \in X'$ and

$$\rho_-(d(x,y)) \le d'(f(x), f(y)) \le \rho_+(d(x,y))$$

for some fixed function $\rho_\pm$ on $[0,\infty)$ with $\lim_{r\to\infty} \rho_-(r) = \infty$. Such f is called a coarse isomorphism. We observe that any two proper left translation invariant metrics d and d' on Γ are *coarsely equivalent* in the sense that the formal identity map from (Γ, d) onto (Γ, d') is a coarse isomorphism. A *coarse metric space* is a space together with a coarse equivalence class of metrics. Hence, Γ is provided with a unique coarse metric space structure. Two groups Γ and Γ' are said to be *coarsely isomorphic* if they are coarsely isomorphic as coarse metric spaces. It follows from the following theorem that exactness is a coarse isomorphism invariant. In particular, a group is exact if it has a finite index subgroup which is exact.

Theorem 2.5 ([47], [83]). *For a group Γ, the following are equivalent.*

1. *The group Γ is exact.*
2. *The metric space (Γ, d) has the property A: For every $\varepsilon > 0$ and $R > 0$, there exist a map $\nu\colon \Gamma \to \mathrm{prob}(\Gamma)$ and $S > 0$ such that $\|\nu_s - \nu_t\| \le \varepsilon$ for every $s, t \in \Gamma$ with $d(s,t) < R$ and $\operatorname{supp} \nu_s \subset \{t : d(s,t) < S\}$ for every $s \in \Gamma$.*
3. *For every $\varepsilon > 0$ and $R > 0$, there exist a Hilbert space $\mathcal{H}$, a map $\xi : \Gamma \to \mathcal{H}$ and $S > 0$ such that $|1 - \langle \xi_t, \xi_s\rangle| < \varepsilon$ for every $s, t \in \Gamma$ with $d(s,t) < R$ and $\langle \xi_t, \xi_s\rangle = 0$ for every $s, t \in \Gamma$ with $d(s,t) \ge S$.*

Moreover, if Γ is exact, then Γ is coarsely isomorphic to a subset of a Hilbert space.

The main result of [83] is the injectivity part of the Baum–Connes conjecture for a group which is coarsely embeddable into a Hilbert space. (See also [45], [76], [84].) This justifies the study of exactness for groups. It is not known whether or not coarse embeddability into a Hilbert space implies exactness (even in the case of groups with the Haagerup property). We recall that a metric space (X, d) has *asymptotic dimension* $\le d$ [36] if for every $R > 0$, there exists a covering $\mathcal{U}$ of X such that $\sup_{U\in\mathcal{U}} \mathrm{diam}(U) < \infty$ and $|\{U \in \mathcal{U} : U \cap B \ne \emptyset\}| \le d + 1$ for any subset $B \subset X$ with $\mathrm{diam}(B) < R$. Asymptotic dimension is a coarse equivalence invariant and hence an invariant for a group. We note that the groups $\mathbb{Z}^d$ and $\mathbb{F}_r^d$ have asymptotic dimension d.

Corollary 2.6 ([47]). *A coarse metric space with finite asymptotic dimension has the property A. In particular, a group with finite asymptotic dimension is exact.*

It was shown in [22] that every Coxeter group has finite asymptotic dimension and hence is exact. We refer to [8] for more information on asymptotic dimension.

We describe a relative version of an amenable action, which is useful in proving various kinds of permanence properties of exactness. There are other approaches [6], [7], [20] which are as well useful. The following is in the spirit of [3].

Proposition 2.7 ([63]). *Let X be a compact Γ-space and K be a countable Γ-space. Assume that there exists a net of Borel maps*

$$\mu_n \colon X \to \mathrm{prob}(K)$$

(i.e., the function $X \ni x \mapsto \mu_n^x(a) \in \mathbb{R}$ is Borel for every $a \in K$) such that

$$\lim_n \int_X \|s.\mu_n^x - \mu_n^{s.x}\|\, dm(x) = 0$$

for every $s \in \Gamma$ and every Radon probability measure m on X. Then Γ is exact provided that all isotropy subgroups of K are exact. Indeed, if Y is a compact Γ-space which is amenable as a Λ-space for every isotropy subgroup Λ, then $X \times Y$ (with the diagonal Γ-action) is an amenable Γ-space.

Corollary 2.8 ([52]). *An extension of exact groups is again exact.*

Proof. If $\Lambda \lhd \Gamma$ is a normal subgroup such that Γ/Λ is exact, then Proposition 2.7 is applicable to an amenable compact (Γ/Λ)-space X and $K = \Gamma/\Lambda$ □

We turn our attention to a group acting on a countable simplicial tree T, which may not be locally finite. We will define a compactification $\overline{T} = T \cup \partial T$ of T, to which Proposition 2.7 is applicable. We recall that a *simplicial tree* is a connected graph without non-trivial circuits, and identify T with its vertex set. The boundary ∂T of T is defined as in Example 2.2. Thus ∂T is the set of all equivalence classes of (one-sided) infinite simple paths in T, where two infinite simple paths are equivalent if their intersection is infinite. For every $a \in T$ and $x \in \partial T$, there exists a unique infinite simple path γ in the equivalence class x which starts at a. We say that the path γ connects a to x. It follows that every two distinct points in $\overline{T} = T \cup \partial T$ are connected by a unique simple path (which is a biinfinite path, with the obvious definition, when both points are boundary points). Every edge separates $\overline{T}$ into two components, and every finite subset of edges separates X into finitely many components. Now we equip $\overline{T}$ with a topology by declaring that all such components are open. It turns out that $\overline{T}$ is compact with this topology. We note that T is dense but not open in $\overline{T}$ (unless T is locally finite) and that every automorphism s of T extends to a

homeomorphism on $\overline{T}$. Fixing a base point $e \in T$, we define $\mu_n \colon \partial T \to \mathrm{prob}(T)$ exactly as in Example 2.2. It is not hard to see that μ_n is a Borel map such that

$$\sup_{x \in \partial T} \|s.\mu_n^x - \mu_n^{s.x}\| \leq \frac{2d(s.e, e)}{n}$$

for every automorphism s of T (cf. Figure 2). We extend μ_n to $\overline{T}$ by simply letting $\mu_n^a = \delta_a \in \mathrm{prob}(T)$ for $a \in T$. Then the sequence of Borel maps $\mu_n \colon \overline{T} \to \mathrm{prob}(T)$ satisfies the assumption of Proposition 2.7 for $X = \overline{T}$, $K = T$ and any group Γ acting on T.

We recall that associated with the fundamental group of a graph of groups there exists a tree, called the *Bass–Serre tree*, on which the group acts. We describe it in the case of an amalgamated free product. Let $\Gamma = \Gamma_1 *_\Lambda \Gamma_2$ be the amalgamated free product of groups Γ_1 and Γ_2 with a common subgroup Λ. Then the associated Bass–Serre tree T is the disjoint union $\Gamma/\Gamma_1 \sqcup \Gamma/\Gamma_2$ of left cosets, where $s\Gamma_1$ and $t\Gamma_2$ are adjacent if $s\Gamma_1 \cap t\Gamma_2 \neq \emptyset$. Thus the edge set of T coincides with Γ/Λ, and an edge $s\Lambda$ connects $s\Gamma_1$ and $s\Gamma_2$. It turns out that T is a tree. The group Γ acts on T from the left in such a way that each vertex stabilizer is conjugate to either Γ_1 or Γ_2 and each edge stabilizer is conjugate to Λ. We note that the tree T is not locally finite unless Λ has finite index in both Γ_1 and Γ_2.

Corollary 2.9 ([25], [78]). *Let Γ be a group acting on a countable simplicial tree T. Then Γ is exact provided that all isotropy subgroups are exact. In particular, an amalgamated free product and an HNN-extension of exact groups are again exact.*

It follows that one-relator groups are exact [38] because they are made up by using HNN-extensions following the McCool–Schupp algorithm. A similar remark applies to a fundamental group of a Haken 3-manifold thanks to the Waldhausen decomposition.

Example 2.2 can be generalized to a hyperbolic space, too. The notion of hyperbolicity was introduced in the very influential paper [35] and has been extensively studied since. A metric space is said to be *hyperbolic* if it is "tree-like" in certain sense, and a finitely generated group Γ is said to be *hyperbolic* if its Cayley graph is hyperbolic. Hyperbolicity is a robust notion and there are many natural examples of hyperbolic groups including the free groups. Every hyperbolic group has a nice compactification, called the Gromov compactification, which is a generalization of that given in Example 2.2. It is shown in [2] that the action of a hyperbolic group on its Gromov compactification is amenable. (See also [9] and the appendix of [5].) The result is generalized in [48], [63] to a group acting on hyperbolic spaces, which are not necessarily locally finite. Compactification of a non-locally-finite hyperbolic graph was considered in [10], where its Bowditch compactification $\overline{K}$ is introduced for a *fine* hyperbolic graph K. A simplicial tree T and its compactification $\overline{T}$ are the simplest non-trivial examples of a uniformly fine hyperbolic graph and its Bowditch

compactification. See [10] for details. As in the case for a simplicial tree, the assumption of Proposition 2.7 is satisfied for a uniformly fine hyperbolic graph K, its Bowditch compactification $\overline{K}$ and any group acting on K [63]. By a characterization of a relatively hyperbolic group [10], we obtain the following corollary.

Corollary 2.10 ([20], [59], [63]). *A relatively hyperbolic group is exact provided that all peripheral subgroups are exact. In particular, every hyperbolic group is exact.*

Examples of relatively hyperbolic groups include the fundamental groups of complete non-compact finite-volume Riemannian manifolds with pinched negative sectional curvature (which are hyperbolic relative to nilpotent cusp subgroups) [26] and limit groups (which are hyperbolic relative to maximal non-cyclic abelian subgroups) [1], [21]. Exactness of limit groups also follows from their linearity.

Another interesting case of group actions which implies exactness is a proper and co-compact action on a finite dimensional CAT(0) cubical complex [12].

The mapping class group $\Gamma(S)$ of a compact orientable surface S is also a natural example of an exact group [42], [49]. Indeed, the action of $\Gamma(S)$ on the space of complete geodesic laminations is amenable [42]. In contrast, the more well-known action of $\Gamma(S)$ on the Thurston boundary $\mathcal{PMF}$ of Teichmüller space is not amenable because of non-amenable isotropy subgroups. However, if we denote by K the set of all non-trivial isotopy classes of non-peripheral simple closed curves on S (i.e., K is the vertex set of the curve complex of S), then the assumption of Proposition 2.7 is satisfied for $X = \mathcal{PMF}$ [49]. Since every isotropy subgroup of a point in K is a mapping class group of lower complexity, induction applies and the exactness of $\Gamma(S)$ follows.

So far we have enumerated examples of exact groups as many as we can (the author is sorry for any possible omission). Unfortunately, there does exist a (finitely presented) group which is neither exact nor coarsely embeddable into a Hilbert space [37]. Currently, it is not known whether the following groups are exact or not: Thompson's group F, $\mathrm{Out}(\mathbb{F}_r)$, automatic groups, 3-manifold groups, groups of homeomorphisms (resp. diffeomorphisms) on (say) the circle S^1, (free) Burnside groups and other monstrous groups.

The rest of this section is devoted to the relationship of exactness to operator algebras. Associated with a group, there are the reduced group C*-algebra $C^*_\lambda(\Gamma)$ and the group von Neumann algebra $\mathcal{L}(\Gamma)$. When Γ is abelian, $C^*_\lambda(\Gamma)$ is isomorphic to $C(\widehat{\Gamma})$, while $\mathcal{L}(\Gamma)$ is isomorphic to $L^\infty(\widehat{\Gamma})$, where $\widehat{\Gamma}$ is the Pontrjagin dual of Γ. Hence the study of $C^*_\lambda(\Gamma)$ corresponds to "noncommutative topology" and that of $\mathcal{L}(\Gamma)$ to "noncommutative measure theory" [15]. Amenability of Γ can be read from its operator algebras.

Theorem 2.11 ([41], [54], [74]). *For a group Γ, the following are equivalent.*

1. *The group Γ is amenable.*
2. *The reduced group C*-algebra $C^*_\lambda(\Gamma)$ is nuclear.*

3. *The group von Neumann algebra $\mathcal{L}(\Gamma)$ is injective.*

A generalization of this theorem to a group action goes as follows.

Theorem 2.12 ([3]). *For a (compact) Γ-space X, the following are equivalent.*

1. *The Γ-space X is amenable.*
2. *The reduced crossed product C*-algebra $C^*_\lambda(X \rtimes \Gamma)$ is nuclear.*
3. *The group-measure-space von Neumann algebra $\mathcal{L}(X \rtimes \Gamma, m)$ is injective for any Γ-quasi-invariant Radon probability measure m on X.*

The nuclear C*-algebras are accessible among the C*-algebras and the classification program of nuclear C*-algebras is a very active area of research in C*-algebra theory [73]. Many C*-algebras $C^*_\lambda(X \rtimes \Gamma)$ arising from various kinds of boundary actions are classifiable via their K-theory [4], [53], [77]. Unlike the group case, a C*-subalgebra of a nuclear C*-algebra needs not be nuclear. The notion of exactness was introduced to give an abstract characterization of subnuclearity and has met a great success [50], [51]. Exactness has a deep connection with operator space theory [51], [69]. A C*-algebra A is called *exact* if taking the minimal tensor product with A preserves short exact sequences of C^*-algebras. The following theorem explains the nomenclature of exact groups.

Theorem 2.13 ([11], [40], [51], [60]). *For a group Γ the following are equivalent.*

1. *The group Γ is exact.*
2. *The reduced group C*-algebra $C^*_\lambda(\Gamma)$ is exact.*
3. *The group von Neumann algebra $\mathcal{L}(\Gamma)$ is weakly exact.*

We note that a C*-subalgebra of an exact C*-algebra is always exact and that a von Neumann subalgebra of a weakly exact von Neumann algebra is weakly exact provided that there exists a normal conditional expectation. Since a von Neumann algebra with a weakly dense exact C*-algebra is weakly exact, we obtain the following corollary.

Corollary 2.14. *Exactness is closed under measure equivalence.*

We recall that two groups Γ and Λ are *measure equivalent* [36] if there exist commuting measure preserving free actions of Γ and Λ on some Lebesgue measure space (Ω, m) such that the action of each of the groups admits a finite measure fundamental domain. For example, lattices in the same (second countable) locally compact group G are measure equivalent. It is known that measure equivalence coincides with the stable orbit equivalence [29] and hence gives rise to a stable isomorphism of the corresponding group-measure-space von Neumann algebras.

3. Amenable compactifications which are small

We study the "size" of an amenable compactification with its application to von Neumann algebra theory in mind. A compactification of a group Γ is a compact space $\Delta\Gamma$ containing Γ as an open dense subset. We only consider those compactifications which are equivariant; the left multiplication action of Γ on Γ extends to a continuous action of Γ on $\Delta\Gamma$. A group Γ is amenable iff the one-point compactification is amenable, and a group Γ is exact iff the Stone–Čech compactification $\beta\Gamma$ is amenable. Thus we think that the "size" of an amenable compactification of a given group measures the "degree of amenability" of the group. We say that a compactification $\Delta\Gamma$ of Γ is *small at infinity* if for every net (s_n) in Γ with $s_n \to x \in \partial\Gamma$, we have $s_n t \to x$ for every $t \in \Gamma$ [13]. In other words, $\Delta\Gamma$ is small at infinity if every flow in Γ drives Γ to a single point. We note that $\Delta\Gamma$ is small at infinity iff the right multiplication action of Γ extends continuously on $\Delta\Gamma$ in such a way that it is trivial on $\Delta\Gamma \setminus \Gamma$. For instance, the Gromov compactification $\mathbb{F}_r \cup \partial\mathbb{F}_r$ of the free group $\mathbb{F}_r$ (cf. Example 2.2) is small at infinity since the first k segment of st is same as that of s as long as $|s| \geq k + |t|$. The same applies to general hyperbolic groups.

We say that a group Γ *belongs to the class* $\mathcal{S}$ if the compact $(\Gamma \times \Gamma)$-space $\partial^\beta\Gamma = \beta\Gamma \setminus \Gamma$ (with the bilateral action) is amenable. If Γ has an amenable compactification $\Delta\Gamma$ which is small at infinity, then we have $\Gamma \in \mathcal{S}$. It follows that the class $\mathcal{S}$ contains amenable groups and hyperbolic groups (or more generally, any group which is hyperbolic relative to a family of amenable subgroups). The class $\mathcal{S}$ is closed under subgroups and free products (with finite amalgamations). Moreover, the wreath product $\Lambda \wr \Gamma$ of an amenable group Λ by a group $\Gamma \in \mathcal{S}$ again belongs to $\mathcal{S}$ [62]. We observe that an inner amenable group in $\mathcal{S}$ has to be amenable because, by definition, a group Γ is *inner amenable* if $\partial^\beta\Gamma$ carries an invariant Radon probability measure for the conjugation action of Γ (cf. [44]). In general, for a given group Γ and a countable Γ-space K, it is an interesting problem to decide whether or not the compact Γ-space $\partial^\beta K = \beta K \setminus K$ is amenable (cf. [62]). We note the trivial case where the isotropy subgroups are all amenable.

The following is a relative version of smallness.

Definition 3.1. Let $\mathcal{G}$ be a non-empty family of subgroups of Γ. For a net (s_n) in Γ we say that $s_n \to \infty$ *relative to* $\mathcal{G}$ if $s_n \notin s\Lambda t$ for any $s, t \in \Gamma$ and $\Lambda \in \mathcal{G}$. We say that a compactification $\Delta\Gamma$ is *small relative to* $\mathcal{G}$ if for every net (s_n) in Γ with $s_n \to x \in \Delta\Gamma$ and with $s_n \to \infty$ relative to $\mathcal{G}$, we have $s_n t \to x$ for every $t \in \Gamma$.

Suppose that a group Γ acts on a simplicial tree T and let $\overline{T}$ be the compactification defined in the previous section. We recall that the open basis of the topology is given by cutting finitely many edges. We fix a base point $e \in T$ and consider the smallest compactification $\Delta^T\Gamma$ of Γ for which the map $\Gamma \ni s \mapsto s.e \in \overline{T}$ is continuous on $\Delta^T\Gamma$. Then $\Delta^T\Gamma$ is small relative to the family of edge stabilizers. Indeed, suppose that $s_n \to \infty$ relative to edge stabilizers and that $a, b \in T$ are given. Let γ be a path connecting a to b. Since the net $(s_n.\gamma)$ of paths leaves every edge, two end points

of $s_n.\gamma$ (i.e., $s_n.a$ and $s_n.b$) converge to the same point (if they converge). The same applies to general fine hyperbolic graphs.

For every non-empty family $\mathcal{G}$ of subgroups of Γ, there exists a compactification $\Delta^{\mathcal{G}}\Gamma$ which is small relative to $\mathcal{G}$ and is largest in the sense that the identity map on Γ extends to a continuous map from $\Delta^{\mathcal{G}}\Gamma$ onto any other compactification which is small relative to $\mathcal{G}$. In the case where $\mathcal{G}$ consists of the trivial subgroup $\{e\}$, a compactification $\Delta\Gamma$ is small relative to $\mathcal{G}$ iff it is small at infinity, and $\Delta^{\mathcal{G}}\Gamma$ is the Higson compactification of the coarse metric space Γ. On the contrary, if $\Gamma \in \mathcal{G}$ then $\Delta^{\mathcal{G}}\Gamma = \beta\Gamma$.

Definition 3.2. Let $\mathcal{G}$ be a non-empty family of subgroups of Γ. We say that $\mathcal{G}$ is *admissible* if there exists an amenable compactification of Γ which is small relative to $\mathcal{G}$, or equivalently if the Γ-space $\Delta^{\mathcal{G}}\Gamma$ is amenable.

From what we have seen, we obtain the following result.

Theorem 3.3. 1. *If Γ acts on a uniformly fine hyperbolic graph K with amenable isotropy subgroups, then the family of edge stabilizers is admissible. In particular, the trivial family of the trivial subgroup $\{e\}$ is admissible for a hyperbolic group.*

2. *Let $\Gamma = \Gamma_1 \times \Gamma_2$ be a direct product and suppose that $\mathcal{G}_i$ are admissible for Γ_i. Then $\mathcal{G} = \{\Gamma_1\} \times \mathcal{G}_2 \cup \mathcal{G}_1 \times \{\Gamma_2\}$ is admissible for Γ.*

3. *Let $\Gamma = \Gamma_1 * \Gamma_2$ be a free product and suppose that $\mathcal{G}_i$ are admissible for Γ_i. Then $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ is admissible for Γ.*

4. Application to von Neumann algebra theory

Let Γ be a group and $\mathbb{C}\Gamma$ be its complex group algebra (with the convolution product). The left regular representation λ of $\mathbb{C}\Gamma$ on $\ell_2(\Gamma)$ is given by

$$(\lambda(f)\xi)(t) = (f * \xi)(t) = \sum_{s\in\Gamma} f(s)\xi(s^{-1}t)$$

for $f \in \mathbb{C}\Gamma$ and $\xi \in \ell_2(\Gamma)$. By taking completion w.r.t. an appropriate topology, we obtain the *group von Neumann algebra* [57]

$$\begin{aligned}\mathcal{L}(\Gamma) &= \text{the weak closure of } \{\lambda(f) : f \in \mathbb{C}\Gamma\} \text{ in } \mathbb{B}(\ell_2(\Gamma))\\ &= \{\lambda(f) : f \text{ a function on } \Gamma \text{ such that } \lambda(f) \text{ is bounded on } \ell_2(\Gamma)\},\end{aligned}$$

where $\mathbb{B}(\ell_2)$ is the algebra of all bounded linear operators on $\ell_2(\Gamma)$. We note that $\mathcal{L}$ is functorial w.r.t. inclusions, direct products and free products. The group von Neumann algebra $\mathcal{L}(\Gamma)$ is *finite* in the sense that it has a faithful finite trace

$$\tau : \mathcal{L}(\Gamma) \ni \lambda(f) \mapsto \langle \lambda(f)\delta_e, \delta_e\rangle = f(e) \in \mathbb{C}.$$

If Γ is an infinite abelian group, then we have $\mathcal{L}(\Gamma) = L^\infty(\widehat{\Gamma}) \cong L^\infty[0,1]$ by uniqueness of the Lebesgue measure space without atoms. (Note that we are still assuming that groups are countable.) Thus, group von Neumann algebras of infinite abelian groups are all isomorphic. The center $Z(\mathcal{L}(\Gamma))$ of $\mathcal{L}(\Gamma)$ is easy to describe;

$$Z(\mathcal{L}(\Gamma)) = \overline{Z(\mathbb{C}\Gamma)}^w = \{\lambda(f) : f \text{ is constant on every conjugacy class}\}.$$

A von Neumann algebra with a trivial center is called a *factor*. Since $f = \lambda(f)\delta_e$ belongs to $\ell_2(\Gamma)$ for every $\lambda(f) \in \mathcal{L}(\Gamma)$, the group von Neumann algebra $\mathcal{L}(\Gamma)$ is a factor iff all non-trivial conjugacy classes of Γ are infinite. Such a group Γ is said to be ICC (abbreviation of "Infinite Conjugacy Classes"). Examples of ICC groups include the free group $\mathbb{F}_r$ and the amenable group $S_\infty = \bigcup S_n$ of finite permutations on a countably infinite set. The classification problem of von Neumann factors was raised in [57], where it is shown that $\mathcal{L}(\mathbb{F}_r) \ncong \mathcal{L}(S_\infty)$. This result is clarified by Theorem 2.11 that Γ is amenable iff $\mathcal{L}(\Gamma)$ is injective. We note that a von Neumann subalgebra of an injective finite von Neumann algebra is again injective and hence injective finite von Neumann algebras are considered "small". Connes's celebrated theorem [14] asserts that $\mathcal{L}(\Gamma) \cong \mathcal{L}(S_\infty)$ for any amenable ICC group Γ. This can be regarded as uniqueness of the amenable noncommutative measure space. In contrast, it is the biggest open problem in the classification of group factors whether $\mathcal{L}(\mathbb{F}_r) \ncong \mathcal{L}(\mathbb{F}_s)$ for $r \neq s$ or not. Free probability theory was invented [80], [82] to tackle this problem and has revealed deep structures of the free group factors [34], [70], [81]. In particular, the free group factors $\mathcal{L}(\mathbb{F}_r)$ $(2 \leq r < \infty)$ are mutually *stably* isomorphic. Moreover, the following dichotomy is known [24], [71]; the free group factors are all isomorphic or all non-isomorphic.

We briefly review the notion of orbit equivalences, which is the ergodic theory counterpart of that of group von Neumann algebras. Let (X, μ) be a Lebesgue probability measure space with a measurable non-singular action of a group Γ. Then we have a group-measure-space von Neumann algebra $\mathcal{L}(X \rtimes \Gamma, \mu)$ which is generated by $L^\infty(X, \mu)$ and a copy of $\mathcal{L}(\Gamma)$ [57]. The von Neumann algebra $\mathcal{L}(X \rtimes \Gamma, \mu)$ is finite if the Γ-action is m.p. (measure preserving) and is a factor if the Γ-action is e.f. (ergodic and free). For two e.f.m.p. actions $\Gamma \curvearrowright (X, \mu)$ and $\Lambda \curvearrowright (Y, \nu)$, we have

$$(L^\infty(X, \mu) \subset \mathcal{L}(X \rtimes \Gamma, \mu)) \cong (L^\infty(Y, \nu) \subset \mathcal{L}(Y \rtimes \Lambda, \nu))$$

iff they are *orbit equivalent* [27], [57], i.e., there exists an isomorphism $F : X \to Y$ of measure spaces such that $F(\Gamma x) = \Lambda F(x)$ for a.e. $x \in X$. Thus the classification of von Neumann algebras and that of orbit equivalences are closely related. We note that it is possible that $\mathcal{L}(X \rtimes \Gamma, \mu) \cong \mathcal{L}(Y \rtimes \Lambda, \nu)$ without being orbit equivalent [17]. We say that two e.f.m.p. actions are *stably orbit equivalent* (or *weakly orbit equivalent*) if they are orbit equivalent "after stabilization", and that two groups Γ and Λ are (resp. *stably*) *orbit equivalent* if they have e.f.m.p. actions which are (resp. stably) orbit equivalent. As we mentioned at the end of Section 2, two groups are stably orbit equivalent iff they are measure equivalent. Connes's aforementioned theorem has the

following counterpart [58]; e.f.m.p. actions of amenable groups are all orbit equivalent to each other. Beyond the amenable case, there has been remarkable progress [86] and exciting new developments [28], [29], [30], [31], [33], [43], [56], [67] in this subject. In particular, it is shown that free groups of different ranks are mutually non-orbit equivalent [30]. We do not further elaborate on ergodic theory, but refer to [32], [68], [75] for details. Before leaving this subject, we mention that as far as we know, the following bold conjecture (communicated to us by D. Shlyakhtenko) stands; ICC groups are (stably) orbit equivalent iff they have (stably) isomorphic group von Neumann algebras.

We now focus on von Neumann algebras. Generally speaking, distinguishing group von Neumann algebras is a difficult task. Indeed, most of known invariants for group von Neumann algebras are binary; injectivity, the property (Γ), the property (T), Haagerup's property, etc. A notable exception is the Cowling–Haagerup constant [19]. Free entropy (dimension) [80], [82] and L^2-homology [18], [55] are candidates for invariants. Recently, a breakthrough was obtained in [66], where a longstanding problem from [57] is solved. It is shown that under certain circumstances, one can specify the position of a prescribed von Neumann subalgebra in the ambient von Neumann algebra. This versatile method found several applications [33], [43], [67], [68]. The following result is obtained by combining this device with theory of exact C*-algebras. In the last few pages, we allow ourselves to be more technical.

Theorem 4.1. *Let Γ be a group and $\mathcal{G}$ be an admissible family of its subgroups. Suppose that $\mathcal{N} \subset \mathcal{L}(\Gamma)$ is an injective von Neumann subalgebra whose relative commutant $\mathcal{N}' \cap \mathcal{L}(\Gamma)$ is non-injective. Then there exist $\Lambda \in \mathcal{G}$ and a non-zero projection $p \in \mathcal{N}$ such that $p\mathcal{N}p$ is conjugated into $\mathcal{L}(\Lambda)$ by a partial isometry in $\mathcal{L}(\Gamma)$.*

We recall that $\mathcal{N}' \cap \mathcal{M} = \{a \in \mathcal{M} : [a, \mathcal{N}] = \{0\}\}$. Sometimes we can patch the pieces $p\mathcal{N}p$ together and find a unitary element $u \in \mathcal{L}(\Gamma)$ such that $u\mathcal{N}u^* \subset \mathcal{L}(\Lambda)$. By Theorems 3.3 and 4.1 and a bit more effort, we obtain the following corollaries.

Recall that a von Neumann algebra $\mathcal{M}$ is *prime* if it does not decompose into a tensor product of two infinite dimensional (diffuse) von Neumann algebras. The free group factor $\mathcal{L}(\mathbb{F}_r)$ is the first example of a separable prime factor [34].

Corollary 4.2 ([61]). *Suppose that Γ belongs to the class $\mathcal{S}$. Then $\mathcal{L}(\Gamma)$ is solid, i.e., for any diffuse subalgebra $\mathcal{N} \subset \mathcal{L}(\Gamma)$, the relative commutant $\mathcal{N}' \cap \mathcal{L}(\Gamma)$ is injective. In particular, $\mathcal{L}(\Gamma)$ is prime unless Γ is amenable.*

Indeed, replacing $\mathcal{N}$ with its maximal abelian subalgebra, we may assume that $\mathcal{N}$ is injective. Then $p\mathcal{N}p$ is conjugated into $\mathcal{L}(\{e\}) = \mathbb{C}$ iff the projection p is atomic in $\mathcal{N}$. Thus, if $\mathcal{N}' \cap \mathcal{L}(\Gamma)$ is non-injective, then $\mathcal{N}$ is not diffuse. We note that if a solid group von Neumann algebra $\mathcal{L}(\Gamma) \cong \mathcal{N}_1 \otimes \mathcal{N}_2$ is not prime, then both $\mathcal{N}_i$ have to be injective and hence $\mathcal{L}(\Gamma)$ itself is injective. Similarly, one can prove that a group-measure-space von Neumann algebra $\mathcal{L}(X \rtimes \Gamma, \mu)$ for $\Gamma \in \mathcal{S}$ is prime [62]. This generalizes a result in [2] that an orbit equivalence relation of a hyperbolic group

is indecomposable. In passing, we mention that an analogous result is obtained for discrete quantum groups and their von Neumann algebras [79]. The following is a von Neumann algebra analogue of the result in [56].

Corollary 4.3 ([64]). *Let $\Gamma_1, \dots, \Gamma_n \in \mathcal{S}$ and assume that they are all non-amenable and ICC. If $\mathcal{M}_1, \dots, \mathcal{M}_m$ are non-injective factors such that*

$$\bigotimes_{j=1}^{m} \mathcal{M}_j \subset \bigotimes_{i=1}^{n} \mathcal{L}(\Gamma_i),$$

then we have $m \leq n$. If in addition $m = n$, then we have "$\mathcal{M}_i \subset \mathcal{L}(\Gamma_i)$" modulo permutation of indices, rescaling, and unitary conjugacy.

We have a Kurosh type theorem for non-prime von Neumann factors. Another version of Kurosh type theorem in presence of rigidity is found in [43], [68].

Corollary 4.4 ([62]). *Let $\Gamma_1, \dots, \Gamma_n$ and $\Lambda_1, \dots, \Lambda_m$ be ICC exact non-amenable groups all of which decompose into non-trivial direct products. Suppose that*

$$\mathcal{L}(\mathbb{F}_\infty * \Lambda_1 * \cdots * \Lambda_m) \cong \mathcal{L}(\mathbb{F}_\infty * \Gamma_1 * \cdots * \Gamma_n).$$

Then $n = m$ and, modulo permutation of indices, $\mathcal{L}(\Lambda_i)$ is unitarily conjugated onto $\mathcal{L}(\Gamma_i)$ for every $i \geq 1$.

It follows that iterated free product factors $\mathcal{L}(\mathbb{F}_\infty * (\mathbb{F}_\infty \times S_\infty)^{*n})$ are mutually non-isomorphic. In contrast, $\mathcal{L}(\mathbb{F}_\infty * (\mathbb{F}_\infty \times \mathbb{Z})^{*n})$ are all isomorphic [23].

We describe one more application of amenable actions in von Neumann algebra theory and ergodic theory. Let Γ be a group acting on a group Λ by automorphisms. Then the semi-direct product $(\Lambda \times \Lambda) \rtimes \Gamma$ naturally acts on Λ, where $\Lambda \times \Lambda$ acts on Λ bilaterally. This action extends to a continuous action on the Stone–Čech compactification $\beta\Lambda$ and then restricts to $\partial^\beta \Lambda = \beta\Lambda \setminus \Lambda$.

Proposition 4.5. *Let Γ and Λ be as above with Λ amenable and assume that the compact $(\Lambda \times \Lambda) \rtimes \Gamma$-space $\partial^\beta \Lambda$ is amenable. Then, for any diffuse von Neumann subalgebra $\mathcal{N} \subset \mathcal{L}(\Lambda) \subset \mathcal{L}(\Lambda \rtimes \Gamma)$, the relative commutant $\mathcal{N}' \cap \mathcal{L}(\Lambda \rtimes \Gamma)$ is injective.*

This proposition applies to the wreath product $\Lambda \wr \Gamma = \left(\bigoplus_\Gamma \Lambda\right) \rtimes \Gamma$ for every amenable group Λ and for every exact group Γ. In the case where Λ is an infinite abelian group, the group factor $\mathcal{L}(\Lambda \wr \Gamma)$ is isomorphic to the group-measure-space factor $\mathcal{L}([0,1]^\Gamma \rtimes \Gamma)$ of the Bernoulli shift action over the base space ([0, 1], Lebesgue). Hence, we obtain the following corollary. We note that the same holds for a noncommutative Bernoulli shift by setting $\Lambda = S_\infty$.

Corollary 4.6 ([62]). *Let Γ be an exact group and $\mathcal{M} = \mathcal{L}([0,1]^\Gamma \rtimes \Gamma)$ be the group-measure-space factor of the Bernoulli shift action. Then, for any diffuse von Neumann subalgebra $\mathcal{A} \subset L^\infty([0,1]^\Gamma)$, the relative commutant $\mathcal{A}' \cap \mathcal{M}$ is injective. In particular, the orbit equivalence relation of a Bernoulli shift action by an exact non-amenable group is indecomposable.*

References

[1] Alibegović, E., A combination theorem for relatively hyperbolic groups. *Bull. London Math. Soc.* **37** (2005), 459–466.

[2] Adams, S., Boundary amenability for word hyperbolic groups and an application to smooth dynamics of simple groups. *Topology* **33** (1994), 765–783.

[3] Anantharaman-Delaroche, C., Systèmes dynamiques non commutatifs et moyennabilité. *Math. Ann.* **279** (1987), 297–315.

[4] Anantharaman-Delaroche, C., Purely infinite C*-algebras arising from dynamical systems. *Bull. Soc. Math. France* **125** (1997), 199–225.

[5] Anantharaman-Delaroche C., Renault, J., *Amenable groupoids*. With a foreword by Georges Skandalis and Appendix B by E. Germain. Monographies de L'Enseignement Mathématique 36, Geneva 2000.

[6] Bell, G. C., Property A for groups acting on metric spaces. *Topology Appl.* **130** (2003), 239–251.

[7] Bell, G., Dranishnikov, A. On asymptotic dimension of groups. *Algebr. Geom. Topol.* **1** (2001), 57–71.

[8] Bell, G., Dranishnikov, A., Asymptotic dimension in Bedlewo. Preprint, 2005.

[9] Biane, P., Germain, E., Actions moyennables et fonctions harmoniques. *C. R. Math. Acad. Sci. Paris* **334** (2002), 355–358.

[10] Bowditch, B. H., Relatively hyperbolic groups. Preprint, 1999.

[11] Brown, N. P., Ozawa, N., *C*-algebras and finite dimensional approximations.* Book in preparation.

[12] Campbell, S. J., Niblo, G. A., Hilbert space compression and exactness for discrete groups. *J. Funct. Anal.* **222** (2005), 292–305.

[13] Carlsson, G., Pedersen, E., Controlled algebra and the Novikov conjectures for K- and L-theory. *Topology* **34** (1995), 731–758.

[14] Connes, A., Classification of injective factors. *Ann. of Math.* (2) **104** (1976), 73–115.

[15] Connes, A., *Noncommutative geometry.* Academic Press, Inc., San Diego, CA, 1994.

[16] Connes, A., Feldman, J., Weiss, B., An amenable equivalence relation is generated by a single transformation. *Ergodic Theory Dynam. Systems* **1** (1981), 431–450.

[17] Connes, A., Jones, V., A II_1 factor with two nonconjugate Cartan subalgebras. *Bull. Amer. Math. Soc.* **6** (1982), 211–212.

[18] Connes, A., Shlyakhtenko, D., L^2-Homology for von Neumann Algebras. Preprint, 2003.

[19] Cowling, M., Haagerup, U., Completely bounded multipliers of the Fourier algebra of a simple Lie group of real rank one. *Invent. Math.* **96** (1989), 507–549.

[20] Dadarlat, M., Guentner, E., Uniform embeddability of relatively hyperbolic groups. Preprint, 2005.

[21] Dahmani, F., Combination of convergence groups. *Geom. Topol.* **7** (2003), 933–963.

[22] Dranishnikov, A., Januszkiewicz, T., Every Coxeter group acts amenably on a compact space. In *Proceedings of the 1999 Topology and Dynamics Conference* (Salt Lake City, UT). Topology Proc. **24** (1999), Spring, 135–141.

[23] Dykema, K., Free products of hyperfinite von Neumann algebras and free dimension. *Duke Math. J.* **69** (1993), 97–119.

[24] Dykema, K., Interpolated free group factors. *Pacific J. Math.* **163** (1994), 123–135.

[25] Dykema, K. J., Exactness of reduced amalgamated free product C^*-algebras. *Forum Math.* **16** (2004), 161–180.

[26] Farb, B., Relatively hyperbolic groups. *Geom. Funct. Anal.* **8** (1998), 810–840.

[27] Feldman, J., Moore, C. C., Ergodic equivalence relations, cohomology, and von Neumann algebras. I, II. *Trans. Amer. Math. Soc.* **234** (1977), 289–359.

[28] Furman, A., Gromov's measure equivalence and rigidity of higher rank lattices. *Ann. of Math.* (2) **150** (1999), 1059–1081.

[29] Furman, A., Orbit equivalence rigidity. *Ann. of Math.* (2) **150** (1999), 1083–1108.

[30] Gaboriau, D., Coût des relations d'équivalence et des groupes. *Invent. Math.* **139** (2000), 41–98.

[31] Gaboriau, D., Invariants l^2 de relations d'équivalence et de groupes. *Inst. Hautes Études Sci. Publ. Math.* **95** (2002), 93–150.

[32] Gaboriau, D., On orbit equivalence of measure preserving actions. In *Rigidity in dynamics and geometry* (Cambridge, 2000), Springer-Verlag, Berlin 2002, 167–186.

[33] Gaboriau, D., Popa, S., An uncountable family of nonorbit equivalent actions of $\mathbb{F}_n$. *J. Amer. Math. Soc.* **18** (2005), 547–559.

[34] Ge, L., Applications of free entropy to finite von Neumann algebras. II. *Ann. of Math.* (2) **147** (1998), 143–157.

[35] Gromov, M., Hyperbolic groups. In *Essays in group theory*, Math. Sci. Res. Inst. Publ., **8**, Springer-Verlag, New York 1987, 75–263.

[36] Gromov, M., Asymptotic invariants of infinite groups. In *Geometric group theory* (Sussex, 1991), Vol. 2, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge 1993, 1–295.

[37] Gromov, M., Random walk in random groups. *Geom. Funct. Anal.* **13** (2003), 73–146.

[38] Guentner, E., Exactness of the one relator groups. *Proc. Amer. Math. Soc.* **130** (2002), 1087–1093.

[39] Guentner, E., Higson N., Weinberger, S., The Novikov Conjecture for Linear Groups. Preprint, 2003.

[40] Guentner, E., Kaminker, J., Exactness and the Novikov conjecture. *Topology* **41** (2002), 411–418.

[41] Hakeda, J., Tomiyama, J., On some extension properties of von Neumann algebras. *Tôhoku Math. J.* (2) **19** (1967) 315–323.

[42] Hamenstädt, U., Geometry of the mapping class groups I: Boundary amenability. Preprint, 2005.

[43] Ioana, A., Peterson, J., Popa, S., Amalgamated free products of w-rigid factors and calculation of their symmetry groups. Preprint, 2005.

[44] de la Harpe, P., Groupes hyperboliques, algèbres d'opérateurs et un théorème de Jolissaint. *C. R. Acad. Sci. Paris Sér. I Math.* **307** (1988), 771–774.

[45] Higson, N., Bivariant K-theory and the Novikov conjecture. *Geom. Funct. Anal.* **10** (2000), 563–581.

[46] Higson, N., Kasparov, G., E-theory and KK-theory for groups which act properly and isometrically on Hilbert space. *Invent. Math.* **144** (2001), 23–74.

[47] Higson N., Roe, J., Amenable group actions and the Novikov conjecture. *J. Reine Angew. Math.* **519** (2000), 143–153.

[48] Kaimanovich, V., Boundary amenability of hyperbolic spaces. In *Discrete geometric analysis*, Contemp. Math. 347, Amer. Math. Soc., Providence, RI, 2004, 83–111.

[49] Kida, Y., The mapping class group from the viewpoint of measure equivalence theory. Preprint, 2005.

[50] Kirchberg, E., On nonsemisplit extensions, tensor products and exactness of group C*-algebras. *Invent. Math.* **112** (1993), 449–489.

[51] Kirchberg, E., Exact C*-algebras, tensor products, and the classification of purely infinite algebras. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 2, Birkhäuser, Basel 1995, 943–954.

[52] Kirchberg, E., Wassermann, S., Permanence properties of C*-exact groups. *Doc. Math.* **4** (1999), 513–558.

[53] Laca, M., Spielberg, J., Purely infinite C*-algebras from boundary actions of discrete groups. *J. Reine Angew. Math.* **480** (1996), 125–139.

[54] Lance, C., On nuclear C*-algebras. *J. Funct. Anal.* **12** (1973), 157–176.

[55] Mineyev, I., Shlyakhtenko, D., Non-microstates free entropy dimension for groups. *Geom. Funct. Anal.* **15** (2005), 476–490.

[56] Monod, N., Shalom, Y., Orbit equivalence rigidity and bounded cohomology. *Ann. of Math.* (2), to appear.

[57] Murray, F. J., von Neumann, J., On rings of operators. IV. *Ann. of Math.* (2) **44** (1943), 716–808.

[58] Ornstein, D. S., Weiss, B., Ergodic theory of amenable group actions. I. The Rohlin lemma. *Bull. Amer. Math. Soc.* **2** (1980), 161–164.

[59] Osin, D. V., Asymptotic dimension of relatively hyperbolic groups. Preprint, 2004.

[60] Ozawa, N., Amenable actions and exactness for discrete groups. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (2000), 691–695.

[61] Ozawa, N., Solid von Neumann algebras. *Acta Math.* **192** (2004), 111–117.

[62] Ozawa, N., A Kurosh type theorem for type II_1 factors. Preprint, 2004.

[63] Ozawa, N., Boundary amenability of relatively hyperbolic groups. *Topology Appl.*, to appear.

[64] Ozawa, N., Popa, S., Some prime factorization results for type II_1 factors. *Invent. Math.* **156** (2004), 223–234.

[65] Paterson, A. L. T., *Amenability*. Math. Surveys Monogr. 29, Amer. Math. Soc., Providence, RI, 1988.

[66] Popa, S., On the fundamental group of type II_1 factors. *Proc. Natl. Acad. Sci. USA* **101** (2004), 723–726.

[67] Popa, S., Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups, I, II. Preprint, 2003.

[68] Popa, S., Deformation and rigidity for group actions and von Neumann algebras. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2006/2007.

[69] Pisier, G., *Introduction to operator space theory.* London Math. Soc. Lecture Note Ser. 294, Cambridge University Press, Cambridge 2003.

[70] Rădulescu, F., The fundamental group of the von Neumann algebra of a free group with infinitely many generators is $\mathbf{R}_+ \setminus \{0\}$. *J. Amer. Math. Soc.* **5** (1992), 517–532.

[71] Rădulescu, F., Random matrices, amalgamated free products and subfactors of the von Neumann algebra of a free group, of noninteger index. *Invent. Math.* **115** (1994), 347–389.

[72] Robertson, G., Steger, T., C^*-algebras arising from group actions on the boundary of a triangle building. *Proc. London Math. Soc.* (3) **72** (1996), 613–637.

[73] Rørdam, M., Structure and classification of C^*-algebras. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1581–1598.

[74] Sakai, S., On the hyperfinite II_1-factor. *Proc. Amer. Math. Soc.* **19** (1968) 589–591.

[75] Shalom, Y., Measurable Group Theory. In *European Congress of Mathematics* (Stockholm, 2004), European Mathematical Society Publishing House, Zürich 2005, 391–424.

[76] Skandalis, G., Tu, J. L., Yu, G., The coarse Baum-Connes conjecture and groupoids. *Topology* **41** (2002), 807–834.

[77] Spielberg, J., Free-product groups, Cuntz-Krieger algebras, and covariant maps. *Internat. J. Math.* **2** (1991), 457–476.

[78] Tu, J. L., Remarks on Yu's "property A" for discrete metric spaces and groups. *Bull. Soc. Math. France* **129** (2001), 115–139.

[79] Vaes, S., Vergnioux, R., The boundary of universal discrete quantum groups, exactness and factoriality. Preprint, 2005.

[80] Voiculescu, D., Free probability theory: random matrices and von Neumann algebras. *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 227–241.

[81] Voiculescu, D., The analogues of entropy and of Fisher's information measure in free probability theory. III. The absence of Cartan subalgebras. *Geom. Funct. Anal.* **6** (1996), 172–199.

[82] Voiculescu, D., Free probability and the von Neumann algebras of free groups. *Rep. Math. Phys.* **55** (2005), 127–133.

[83] Yu, G., The coarse Baum-Connes conjecture for spaces which admit a uniform embedding into Hilbert space. *Invent. Math.* **139** (2000), 201–240.

[84] Yu, G., Higher index theory of elliptic operators and geometry of groups. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1623–1639.

[85] Zimmer, R. J., Amenable ergodic group actions and an application to Poisson boundaries of random walks. J. Funct. Anal. **27** (1978), 350–372.

[86] Zimmer, R. J., *Ergodic theory and semisimple groups.* Monogr. Math. 81, Birkhäuser, Basel 1984.

Department of Mathematical Sciences, University of Tokyo, Komaba, 153-8914, Japan
and
Department of Mathematics, UCLA, Los Angeles, CA 90095, U.S.A.
E-mail: narutaka@ms.u-tokyo.ac.jp

Structure and classification of C^*-algebras

Mikael Rørdam

Abstract. We give an overview of the development over the last 15 years of the theory of simple C^*-algebras, in particular in regards to their classification and structure. We discuss dimension theory for (simple) C^*-algebras, in particular the so-called stable and real ranks, and we explain how properties of C^*-algebras of low dimension (stable rank one and real rank zero) was used by the author and P. Friis to give a new and simple proof of a theorem of H. Lin that almost commuting self-adjoint matrices are close to exactly commuting self-adjoint matrices. Elliott's classification program is explained and is contrasted with recent examples of C^*-algebras of "high dimension", including an example of a simple C^*-algebra with a finite and an infinite projection.

Mathematics Subject Classification (2000). Primary 46L35; Secondary 46L80.

Keywords. Simple C^*-algebras, classification, K-theory, dimension, almost commuting matrices.

1. Introduction

A (represented) C^*-algebra is a norm closed self-adjoint sub-algebra of the bounded operators on a Hilbert space. One can alternatively describe C^*-algebras axiomatically as complex Banach algebras with an involution that satisfies $\|a^*a\| = \|a\|^2$, thanks to a theorem of Gelfand, Naimark and Segal. Each commutative C^*-algebra is isomorphic to $C_0(X)$ for some locally compact Hausdorff space X, and $C_0(X)$ is isomorphic to $C_0(Y)$ if and only if X and Y are homeomorphic. This justifies the jargon that the study of C^*-algebras is *non-commutative topology*.

One can associate a C^*-algebra to each (locally compact) group, and the representation theory of the group C^*-algebra coincides with the representation theory of the group. Much of the early interest in C^*-algebras lay in their representation theory, and in their connection with other objects (such as groups).

Nearly 50 years ago, Glimm constructed a class of C^*-algebras, now called UHF- or Glimm algebras, that are the C^*-algebra analog of the hyperfinite II_1-factor. Unlike the situation for von Neumann algebras, there is not one UHF-algebra but in fact uncountably many. Glimm classified UHF-algebras by an invariant that we today can identify as the K_0-group of the algebra. Glimm's work was extended by Bratteli and Elliott who classified the larger class of AF-algebras (approximately finite dimensional C^*-algebras) that arise as inductive limits of finite dimensional C^*-algebras (the latter

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

are just direct sums of matrix algebras). All UHF-algebras are *simple*, ie. have no non-trivial closed two-sided ideals. AF-algebras may or may not be simple, and far from all simple AF-algebras are UHF-algebras. Simple AF-algebras need not have unique trace, actually any metrizable Choquet simplex can arise as the trace simplex of a simple unital AF-algebra.

Many other interesting examples of (simple) C^*-algebras saw the light in the 1970s and 1980s. Cuntz invented his algebras $\mathcal{O}_n$. These are, for $2 \leq n < \infty$, generated by n isometries $s_1, s_2, \ldots, s_n$ satisfying the relation $1 = s_1 s_1^* + s_2 s_2^* + \cdots + s_n s_n^*$. (For $n = \infty$ this relation is replaced with the relation that the support projections $s_j s_j^*$ are mutually orthogonal.) Cuntz proved that his algebras are simple and purely infinite (see Definition 2.3) and independent on the choice of generators. These were the first explicit examples of simple infinite separable C^*-algebras. Cuntz and Krieger later associated a C^*-algebra to each finite Markov chain. This construction has today been generalized considerably in parts by Pimsner, who associated a Pimsner algebra to each Hilbert bi-module over a C^*-algebra, and with the constructions of C^*-algebras associated with infinite graphs. Many of these C^*-algebras are simple and purely infinite.

One can also obtain simple C^*-algebras from groups. The reduced groups C^*-algebra $C^*_{\mathrm{red}}(G)$ associated with a (discrete) group is simple for many interesting cases of non-amenable groups G, for example when G is a free group (other than $\mathbb{Z}$). These C^*-algebras are often exact, but never nuclear. Non-amenable groups can act amenably on spaces and can in this way give rise to simple, purely infinite, and nuclear C^*-algebra. Dynamical systems in general, also with amenable groups and in particular with $\mathbb{Z}$, give rise to many interesting examples of C^*-algebras, many of which are simple.

The irrational rotation C^*-algebra, A_θ, associated with an irrational number θ, is the universal C^*-algebra generated by two unitaries u and v satisfying the commutation relation $uv = e^{2\pi i\theta} vu$. They were first studied by Rieffel and shown to be simple with a unique trace and being independent of the generators u and v. The irrational rotation C^*-algebra A_θ contains the Harper operator $u + u^* + \lambda(v + v^*)$, where λ is a non-zero real parameter, whose spectrum recently has been shown to be a Cantor set.

These examples, and many more like them, have spurred the interest in understanding, and perhaps classifying, C^*-algebras, in particular the simple ones. This study was first suggested by Dixmier in the 1960s, and later taken up by Cuntz and Blackadar to mention just a few. It was investigated when finite simple C^*-algebras have a trace, and Cuntz studied the purely infinite C^*-algebras (that resemble the type III_1 von Neumann factors). The question, if all simple C^*-algebras are either (stably) finite or purely infinite was left open until a few years ago where the author found a counterexample inspired by ideas of Villadsen.

The most significant progress in our understanding of C^*-algebras comes from the program initiated by Elliott, and known as Elliott's classification program. Elliott predicts that (simple) separable nuclear C^*-algebras can be classified by natural invariants including K-theory as the most prominent ingredient. This conjecture

has now been verified for a surprisingly wide class of C^*-algebras, for example for all simple separable nuclear purely infinite K-amenable C^*-algebras (the Kirchberg–Phillips theorem). We also know that we must make modifications to the classification conjecture if we want to turn it into a theorem.

We give here an overview of the theory of simple C^*-algebras including some of the recent examples of exotic "high-dimensional" simple C^*-algebras. We also include a solution to a classical problem, if almost commuting matrices must be close to commuting matrices, as the methods to solve this problem grew out of the methods used to study (simple) C^*-algebras.

2. The structure of simple C^*-algebras

Von Neumann algebra factors were by their inventors, von Neumann and Murray, divided into types: I_n, I_∞, II_1, II_∞, and III. The types are distinguished by the *dimension range* of the projections in the factor, which for the 5 types above are $\{0, 1, 2, \ldots, n\}$, $\{0, 1, 2, \ldots, \infty\}$, $[0, 1]$, $[0, \infty]$, and $\{0, \infty\}$, respectively. A type I_n-factor, with n finite, is isomorphic to the algebra of $n \times n$ matrices, and, more generally a type I_n factor is the algebra of all bounded operators on an n-dimensional Hilbert space. Type II_1-factors admit a unique tracial state, and type III-factors are traceless. A separable von Neumann algebra is simple (has no non-trivial closed two-sided ideals) if and only if it is a factor of type I_n, with n finite, type II_1, or of type III.

Can one similarly divide the (infinite dimensional) simple C^*-algebras into two types; a *finite type* resembling the type II_1-factors and an *infinite type* resembling the type III-factors? Existence of traces and of finite and infinite projections should be natural dividing criterions:

Definition 2.1. Two projections p and q in a C^*-algebra A are said to be (Murray–von Neumann) *equivalent*, written $p \sim q$, if $p = v^*v$ and $q = vv^*$ for some (partial isometry) v in A; and p is *subequivalent* to q, written $p \precsim q$, if p is equivalent to a subprojection of q.

A projection in a C^*-algebra A is said to be *infinite* if it is equivalent to a proper subprojection of itself; and it is said to be *finite* otherwise.

A simple C^*-algebra A is called *stably infinite* if its stabilization $A \otimes \mathcal{K}$ contains an infinite projection, and it is called *stably finite* otherwise.

The notion of finiteness relate, as we would expect, to the existence of traces. As C^*-algebras need not be unital, we allow our traces to be unbounded and densely (not necessarily everywhere) defined.

The usual construction of a trace on an abstract C^*-algebra goes via a so-called *dimension function* (a "measure" rather than the "integral"), which by "integration" gives rise to a functional, which is slightly short of being a trace: additivity is known to hold only on abelian subalgebras. Such functionals are called *quasitraces*. Uffe

Haagerup proved in [16] that quasitraces are in fact traces on *exact* C^*-algebras (Haagerup proved this for unital C^*-algebras, and Kirchberg extended the result to the non-unital case).

We have the following fundamental theorem on the existence of traces on simple C^*-algebra:

Theorem 2.2 (Blackadar-Cuntz–Haagerup). *A simple C^*-algebra A admits a quasitrace* (*and hence a trace, if A is exact*) *if and only if A is stably finite.*

Outline of proof: Blackadar and Cuntz proved in [2] that the following three conditions are equivalent for a simple *stable* C^*-algebra A: 1) A contains an infinite projection, 2) A has no dimension function, and 3) A is algebraically simple. Any dimension function lifts to a quasitrace by [3], so the equivalence of 1) and 2) together with Haagerup's result, that quasitraces on exact C^*-algebras are traces, yields the theorem.

How finite are stably finite simple C^*-algebras? and how infinite are the stably infinite ones? The definition below, due to Cuntz, is relevant for the discussion of the latter.

Definition 2.3. A simple C^*-algebra A is said to be *purely infinite* if every non-zero hereditary subalgebra of A contains an infinite projection.

Any subalgebra of the form $\overline{xAx^*}$ is hereditary, and the converse holds in the separable case. In other words, a simple C^*-algebra is purely infinite if one can find infinite projections in all "arbitrarily small corners" of A. A purely infinite C^*-algebra is clearly stably infinite. The opposite does not hold as we shall see in Section 5.

2.1. Dimensions of C^*-algebras. A commutative C^*-algebra is isomorphic to $C_0(X)$ for some locally compact Hausdorff space X, and the space X is determined up to homeomorphism by the isomorphism class of the C^*-algebra. In the commutative case we can therefore define the dimension of the C^*-algebra to be the classical dimension of the space X. What about the non-commutative case? It turns out that there are several, and unfortunately mutually disagreeing, ways of extending dimension to the non-commutative setting. The low dimension cases are of most interest in particular in the study of simple C^*-algebras (many nicely behaving simple C^*-algebras are of very low dimension). Two notions of "low dimension" are particularly important:

Definition 2.4. Let A be a C^*-algebra. If the set of invertible elements in A (or in the unitization of A, if A is non-unital) is dense in A, then A is said to be of *stable rank one*, written $\mathrm{sr}(A) = 1$.

If the set of *self-adjoint* invertible elements in A (or in the unitization of A, if A is non-unital) is dense in the set of self-adjoint elements in A, then A is said to be of *real rank zero*, written $\mathrm{RR}(A) = 0$.

Rieffel introduced stable rank in his paper [24], and Brown and Pedersen introduced real rank in [5]. A commutative C^*-algebra $C_0(X)$ is of stable rank one if $\dim(X) \leq 1$, and of real rank zero if $\dim(X) = 0$.

Brown and Pedersen show that a C^*-algebra is of real rank zero if and only if the set of self-adjoint elements with finite spectrum is dense in the set of all self-adjoint elements. All purely infinite simple C^*-algebras are of real rank zero:

Proposition 2.5 (Cuntz [6], Zhang [35]). *The following three conditions are equivalent for any simple C^*-algebra (other than $\mathbb{C}$):*

(i) *A is purely infinite,*

(ii) *for all non-zero positive elements a, b in A there exists $x \in A$ such that $b = x^*ax$,*

(iii) $\mathrm{RR}(A) = 0$ *and all non-zero projections in A are infinite.*

Stably infinite C^*-algebras are never of stable rank one (in fact they have stable rank $+\infty$). It was a surprise when Villadsen in [33] showed that stably finite simple C^*-algebras need not be of stable rank one. Stably finite C^*-algebras can have very few projections and hence have real rank greater than zero.

2.2. Comparison theory for C^*-algebras. Comparison theory for projections in von Neumann algebras is a crucial ingredient in the classification of von Neumann factors into types and to proving existence of traces on finite von Neumann algebras. Comparison of projections in a von Neumann factor is total: for any two projections p, q one has either $p \precsim q$ or $q \precsim p$ (see Definition 2.1). The comparison theory for C^*-algebras is far more delicate as is in parts reflected by looking at the ordered K_0-group. Any simple dimension group arises as the ordered K_0-group of a simple AF-algebra, and such ordered groups easily fail to be totally ordered. The second best thing after total comparison of projection is weak (or almost) unperforation, described below.

The comparison properties for a C^*-algebra A are contained in the ordered monoids $V(A)$ and $W(A)$ consisting of equivalence classes of projections and positive elements, respectively, in the (non-unital) *-algebra $M_\infty(A) = \bigcup_{n=1}^{\infty} M_n(A)$. Equivalence of projections is the usual Murray–von Neumann equivalence (see Definition 2.1). Following Cuntz, comparison of positive elements $a, b \in M_\infty(A)$ is defined as follows: $a \precsim b$ if there is a sequence $\{x_n\}$ in $M_\infty(A)$ such that $x_n^* b x_n \to a$; and by $a \approx b$ iff $a \precsim b$ and $b \precsim a$ one defines an equivalence relation on the positive elements, which by the way does not quite agree with Murray–von Neumann equivalence when restricted to projections.

The sets $V(A)$ and $W(A)$ become ordered abelian semigroups by defining addition to be "orthogonal addition":

$$[a] + [b] = \left[\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \right],$$

and ordering to be induced by $\precsim$. The ordering on $V(A)$ coincides with the algebraic ordering: $x \leq y$ iff there is z such that $y = x + z$. The ordering on $W(A)$ is not

algebraic. Both semigroups are *positive* in the sense that they have a zero element which at the same time is the smallest element of the semigroup; hence $x \leq x + y$ for all x, y. The semigroup $V(A)$ is called the Murray–von Neumann semigroup of A, and $W(A)$ is called the Cuntz semigroup of A.

If A is generated as an ideal by its projections (which is the case for all simple C^*-algebras with a non-trivial projection), then $K_0(A)$ is the Grothendieck group of $V(A)$, and the positive cone, $K_0(A)^+$, is the image of $V(A)$ in $K_0(A)$.

It was shown in [3] that there is a one-to-one correspondence between (lower semi-continuous) states on $W(A)$ and quasitraces on A, and by Haagerup's theorem in [16], quasitraces are traces on exact C^*-algebras. States on $V(A)$ extends (possibly non-uniquely) to (lower semi-continuous) states on $W(A)$ as shown in [4]. It follows in particular that each state on $V(A)$ lifts to a trace on A if A is exact. A new proof of this fact was recently given by Haagerup and Thorbjørnsen, [17], using random matrix methods.

An ordered abelian positive semigroup $(W, +, \leq)$ is said to be *almost unperforated* if

$$\forall n, m \in \mathbb{N}\ \forall x, y \in W : nx \leq my \text{ and } n > m \implies x \leq y.$$

(The negation of almost unperforation is *strong perforation*.) One can use a Hahn–Banach type argument (see [15] and [29]) to show that $(W, +, \leq)$ is almost unperforated if and only if the order on W is determined by states on W. It follows in particular, that if A is simple and exact, if $V(A)$ is almost unperforated, and if p, q are two projections in $M_\infty(A)$, then $p \precsim q$ if $\tau(p) < \tau(q)$ for all traces τ on A. A similar statement, with dimension functions in the place of traces, holds for $W(A)$ (see [29]).

A simple C^*-algebra A is purely infinite if and only if $W(A)$ has only one non-zero element; and if $W(A)$ is almost unperforated, then A is either purely infinite or stably finite. It is known that $W(A)$ and $V(A)$ are almost unperforated for many C^*-algebras of interest including, besides all purely infinite C^*-algebras, also all C^*-algebras that tensorially absorb the Jiang-Su algebra $\mathcal{Z}$ (see [29]).

It is quite often the case that the semigroups $V(A)$ and $W(A)$ are almost unperforated, but it is not true in general for simple C^*-algebras as shown in the pioneering work of Villadsen (see Section 5). Almost unperforation can fail spectacularly. For example there is a simple nuclear C^*-algebra A in which one has elements $x, y_1, y_2, y_3, \ldots \in V(A)$ satisfying $2x = 2y_1 = 2y_2 = \cdots$ and $x \not\leq y_1 + y_2 + \cdots + y_n$ for all natural numbers n, see [26].

It is not known if such exotic phenomenons can occur for C^*-algebras of real rank zero:

Question 2.6. Suppose that A is a simple C^*-algebra of real rank zero.

(i) Does it follow that A is either stably finite or purely infinite?

(ii) Does it follow that $V(A)$ and $W(A)$ are almost unperforated?

2.3. Tensor products and free products. Takesaki proved that the minimal (= spatial) tensor product of two simple C^*-algebras is again simple. This is at first thought perhaps not surprising, but one should bear in mind that the minimal tensor product of two (non-simple and non-exact) C^*-algebras can have unexpected and exotic ideals.

Following the similar notion from von Neumann factors we say that a simple C^*-algebra is *tensorially prime* if it is not isomorphic to a tensor product $A \otimes B$, where both A and B are (simple and) non-type I (i.e., are not isomorphic to the compact operators on a finite or infinite dimensional Hilbert space). We consider here only the minimal tensor product, which we denote by $\otimes$.

Proposition 2.7 (Kirchberg, see [27]). *Let A and B be simple non-type I C^*-algebras. If A or B is stably infinite, then $A \otimes B$ is purely infinite. If A and B are both stably finite* and exact*, then $A \otimes B$ is stably finite.*

In particular, if D is a simple, exact, and non-tensorially prime C^*-algebra, then D is either stably finite and admits a trace or D is purely infinite.

Note that we do not know if the tensor product of two (non-exact) stably finite simple C^*-algebras is stably finite. This would be the case if we knew that quasitraces on arbitrary (non-exact) C^*-algebras are traces.

Several simple C^*-algebras are non-tensorially prime without obviously being so. Jiang and Su constructed in [18] a simple separable unital stably finite non-type I C^*-algebra $\mathcal{Z}$, called the Jiang–Su algebra, that has the same K-theory (and the same Elliott invariant, see Section 3) as the complex numbers. It has been shown that many C^*-algebras are $\mathcal{Z}$-absorbing, i.e., they satisfy $A \cong A \otimes \mathcal{Z}$; $\mathcal{Z}$-absorbing C^*-algebras are obviously non-tensorially prime.

The most non-commutative product of two C^*-algebras is the free product (= the "largest" C^*-algebra generated by copies of the two C^*-algebras). We also have the unital free product $A *_{\mathbb{C}} B$ of two unital C^*-algebras A and B, which is defined to be the "largest" unital C^*-algebras generated by a unital copy of A and a unital copy of B.

Consider the C^*-algebra $A = M_2 *_{\mathbb{C}} \mathcal{O}_2$, and let $e \in M_2 \subseteq A$ be a 1-dimensional projection in M_2. Then $e \oplus e$ is equivalent to 1_A, which is a (properly) infinite projection in A. The projection e is finite in A, intuitively because it is in free position from $\mathcal{O}_2$ (a rigorous proof of this fact is non-trivial). The free product C^*-algebra A however is very far away from being simple.

Voiculescu introduced the notion of *reduced free products* of C^*-algebras, or rather of non-commutative probability spaces (A, ρ_A) and (B, ρ_B), where A and B are unital C^*-algebras, and ρ_A and ρ_B are states on A and B, respectively. The reduced free product is again a non-commutative probability space, denoted $(A *_{\mathrm{red}} B, \rho_A * \rho_B)$. The associated C^*-algebra $A *_{\mathrm{red}} B$ is very often simple (see [1]), and it is exact if both A and B are exact (see [8] and [11]), but it is almost never nuclear.

At a first glance one might expect that the (simple) reduced free product C^*-algebra $M_2 *_{\mathrm{red}} \mathcal{O}_2$ (with respect to suitable states on M_2 and $\mathcal{O}_2$) would be an example of an infinite C^*-algebra with a finite projection $e \in M_2$ (as above). However, it

turns out that most reduced free product C^*-algebras, including $M_2 *_{\mathrm{red}} \mathcal{O}_2$, have excellent comparison theory (eg., their Murray–von Neumann semigroup is almost unperforated), and one can show that the projection e from above is infinite in the reduced free product, and, moreover, that $M_2 *_{\mathrm{red}} \mathcal{O}_2$ (and other C^*-algebras like it) is purely infinite. Results along these lines were obtained by Dykema and the author in [9] and [10].

3. Elliott's classification conjecture

The possibility that C^*-algebras can be classified – up to *-isomorphism – by K-theory was perhaps first suggested by Glimm's classification of UHF-algebras (also called Glimm algebras) by supernatural numbers, or, equivalently, by a subgroup of the rational numbers, their K_0-group. This classification was later extended to AF-algebras by Bratteli and Elliott to yield a one-to-one correspondence between dimension groups and AF-algebras. The former were axiomatically described by Effros, Handelman, and Shen as being the unperforated ordered abelian groups with the Riesz Interpolation Property. In the late 1980s in [12] Elliott extended the classification of AF-algebras to include a class of C^*-algebras, now called A$\mathbb{T}$-algebras, that arise as inductive limits of direct sums of matrix algebras over $C(\mathbb{T})$, with the added assumption that the inductive limit C^*-algebra is of real rank zero. These algebras can have non-trivial K_1-group. Elliott raised in the same paper the possibility that his classification might encompass much more than this apparently rather special class of C^*-algebras: many naturally occurring C^*-algebras might be A$\mathbb{T}$-algebras of real rank zero. Moreover, the same invariant, or possibly an expanded version of it, might classify an even wider class of C^*-algebras. These ideas, expressed in more detail below, are known as the *Elliott classification conjecture*.

Elliott's prediction, that A$\mathbb{T}$-algebras of real rank zero are rather frequently occurring C^*-algebras, was shortly after confirmed by himself and Evans as they discovered that the irrational rotation C^*-algebras mentioned in the introduction are A$\mathbb{T}$-algebras. Putnam showed around the same time that C^*-algebras associated with a minimal action on the Cantor set likewise are A$\mathbb{T}$-algebras.

Turning to the precise formulation of the classification conjecture, we only expect to be able to deal with separable and nuclear C^*-algebras (nuclearity is for C^*-algebras what injectivity, or equivalently, hyperfiniteness, is for von Neumann algebras). The K-theory of a C^*-algebra A consists of two abelian groups $K_0(A)$ and $K_1(A)$. The K_0-group has a distinguished subset, $K_0(A)^+$, (the image of $V(A)$ in $K_0(A)$, cf. Section 2), which gives $K_0(A)$ the structure of an ordered abelian group when A has an approximate unit consisting of projections and when A is stably finite.

To simplify its statement, and to state the conjecture in a situation, where no counterexamples (yet) exist, we state formally the Elliott conjecture only in the real rank zero case:

Conjecture 3.1 (Elliott – the real rank zero case). Let A and B be simple separable nuclear unital C^*-algebras of real rank zero. Then

$$A \cong B \iff (K_0(A), K_0(A)^+, [1_A], K_1(A)) \cong (K_0(B), K_0(B)^+, [1_B], K_1(B)).$$

The isomorphism on the right-hand side asserts that there exist isomorphisms $\alpha_0 \colon K_0(A) \to K_0(B)$ and $\alpha_1 \colon K_1(A) \to K_1(B)$ such that $\alpha_0(K_0(A)^+) = K_0(B)^+$ and $\alpha_0([1_A]) = [1_B]$. The invariant can detect whether A is stably finite or stably infinite: $K_0(A)^+ = K_0(A)$ in the latter case, and $K_0(A) \neq 0$ and $K_0(A) \cap -K_0(A)^+ = 0$ in the former case.

The conjecture can – with due care – be extended to non-simple C^*-algebras. We have already mentioned that Elliott's results in [12] confirms his conjecture for A$\mathbb{T}$-algebras of real rank zero. Dadarlat and Gong, [7], later verified the conjecture for the much wider class of so-called AH-algebras (of slow dimension growth) of real rank zero. These classification results also hold in the non-simple case, but the invariant becomes more complicated. It is an open problem if all simple separable nuclear stably finite C^*-algebras of real rank zero are AH-algebras of slow dimension growth and hence classifiable. The range of the invariant has been completely described by Elliott and Gong (see [27, Proposition 3.3]).

K-theory alone will not classify stably finite C^*-algebras not of real rank zero. Intuitively, if a C^*-algebra has very few – or no – projections, then its K_0-group probably say less about the algebra, so we need more information in our invariant. Goodearl produced a class of C^*-algebras (now known as Goodearl algebras) where the trace simplex of the C^*-algebra cannot be detected from its K-theory. This suggests that the trace simplex must be included in the invariant, and – as pointed out by Thomsen – also the pairing between traces and K_0. The resulting invariant (see eg. [27, Chapter 3]) is known as the *Elliott invariant*. The literature contains strong classification results in terms this invariant also for non-real rank zero C^*-algebras, eg. the classification of all simple AH-algebras of bounded dimension by Elliott, Gong and Li, [13], and there is a good description of the range of the invariant for this class due to Villadsen (see [27, Proposition 3.3.7]). A more ultimate result on the range of the invariant within the still not classified class of so-called ASH-algebras due to Elliott and Thomsen can be looked up in [27, Theorem 3.4.4].

The best classification results exist in the stably infinite case. There are no traces on simple stably infinite C^*-algebras, and the order structure on K_0 degenerates: $K_0^+ = K_0$. The Elliott invariant therefore collapses to the two groups $K_0(A)$ and $K_1(A)$ with no other structure except the position of the unit in $K_0(A)$ in the unital case.

The classification result below, that confirms the Elliott conjecture for a sweeping class of stably infinite C^*-algebras, was obtained independently by Kirchberg and Phillips, [19] and [23]:

Theorem 3.2 (Kirchberg–Phillips). *Let A and B be separable, nuclear, simple, purely infinite, K-amenable, unital C^*-algebras. Then*

$$A \cong B \iff (K_0(A), [1_A], K_1(A)) \cong (K_0(B), [1_B], K_1(B)).$$

A C^*-algebra A is *K-amenable* if it is KK-equivalent to an abelian C^*-algebra; and the class of K-amenable C^*-algebras forms a bootstrap class, see [30]. Two K-amenable C^*-algebras are KK-equivalent if and only if they have isomorphic K-groups. One can remove the condition that A and B are K-amenable by replacing the assumption that the K-groups are isomorphic with the assumption that A and B are KK-equivalent. It is an important open problem if all nuclear C^*-algebras are K-amenable.

The Kirchberg–Philips theorem verifies the Elliott conjecture in the stably infinite, real rank zero case modulo two open problems: Are all separable simple nuclear stably infinite C^*-algebras of real rank zero purely infinite (cf. Question 2.6 (i))? And the problem above if all (separable, simple, purely infinite) nuclear C^*-algebras are K-amenable.

The range of the invariant in the stably infinite case is easy to describe: all pairs of countable abelian groups can arise as K_0 and K_1, and there are no restriction on the position of the unit, see [27, Propositions 4.3.3 and 4.3.4]. The Elliott conjecture would predict that all separable nuclear simple stably infinite C^*-algebras are actually purely infinite. As already mentioned, and as will be shown in Section 5, this is not the case. It may still be that separable nuclear simple stably infinite C^*-algebras of *real rank zero* are purely infinite, cf. Question 2.6 and that the Elliott conjecture holds for these C^*-algebras.

The status for the Elliott conjecture is nonetheless open. It may be that the invariant will be refined, so that it can distinguish also the "high-dimensional" examples that we shall discuss in Section 5, but it may also be that the class of C^*-algebras to be classified must be restricted, for example to the class of $\mathcal{Z}$-absorbing C^*-algebras (that briefly were discussed at the end of Section 2). There are some positive results in this direction, eg. by W. Winter, [34], who verified Elliott's conjecture for $\mathcal{Z}$-stable C^*-algebras of real rank zero and with finite decomposition rank.

4. Almost commuting self-adjoint matrices: an application of real rank zero and stable rank one

The classical problem, if two almost commuting self-adjoint matrices are close to two exactly commuting self-adjoint matrices, was solved in the early 1990s by Huaxin Lin, [20], using techniques from C^*-algebras. His long and technical proof was shortened significantly by Friis and the author, [14], where the analysis was reduced to using known properties of C^*-algebras of real rank zero and stable rank one. We outline the ideas of this argument here, and begin by stating the exact formulation of Lin's theorem:

Theorem 4.1. *For each $\varepsilon > 0$ there is a $\delta > 0$ such that for every natural number n and for every pair of self-adjoint $n \times n$ matrices A and B satisfying*

$$\|AB - BA\| < \delta, \quad \|A\| \leq 1, \ \|B\| \leq 1,$$

there exists a pair of commuting self-adjoint $n \times n$ matrices A' and B' such that $\|A - A'\| \leq \varepsilon$ and $\|B - B'\| \leq \varepsilon$.

As an instructive example of almost commuting self-adjoint matrices that not obviously are close to commuting self-adjoint matrices, consider the following $n \times n$ matrices:

$$A_n = \begin{pmatrix} 1/n & 0 & 0 & \cdots & 0 \\ 0 & 2/n & 0 & \cdots & 0 \\ 0 & 0 & 3/n & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad B_n = \begin{pmatrix} 0 & 1/2 & 0 & \cdots & 0 \\ 1/2 & 0 & 1/2 & \cdots & 0 \\ 0 & 1/2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Note that $\|A_n B_n - B_n A_n\| \leq 1/n \to 0$ as $n \to \infty$. It follows from Theorem 4.1 that there are commuting $n \times n$ matrices A'_n and B'_n such that $\|A_n - A'_n\| \to 0$ and $\|B_n - B'_n\| \to 0$. Curiously, there are – to the knowledge of the author – no known *explicit* choices for such sequences of matrices $\{A'_n\}$ and $\{B'_n\}$.

The theorem is proved indirectly. If it were wrong, then there would exist a counterexample: $\varepsilon > 0$ and sequences $\{A_n\}$ and $\{B_n\}$ of self-adjoint $k_n \times k_n$ matrices all of norm at most one such that $\|A_n B_n - B_n A_n\| \to 0$ and such that the distance from (A_n, B_n) to a commuting pair of self-adjoint matrices is at least ε for all n. We show that the existence of such a counterexample leads to a contradiction.

Set $T_n = A_n + i B_n$, and note that $\|T_n\| \leq 2$ and that $\|T_n T_n^* - T_n^* T_n\| \to 0$. Let $\mathfrak{A} = \prod_{n=1}^{\infty} M_{k_n}$ be the ℓ^∞-direct product of the matrix algebras and let $\mathfrak{I} = \sum_{n=1}^{\infty} M_{k_n}$ be the c_0-direct sum of matrix algebras. Then $\mathfrak{I}$ is a closed two-sided ideal in $\mathfrak{A}$, and so we can consider the quotient $\mathfrak{B} = \mathfrak{A}/\mathfrak{I}$ and the quotient mapping $\pi : \mathfrak{A} \to \mathfrak{B}$. Put $T = \{T_n\} \in \mathfrak{A}$. Then $TT^* - T^*T$ belongs to $\mathfrak{I}$, and so $\pi(T)$ is a normal operator in the C^*-algebra $\mathfrak{B}$.

If we could lift $\pi(T)$ to a normal operator $S = \{S_n\}$ in $\mathfrak{A}$, then we would have our contradiction: Write $S_n = A'_n + i B'_n$, with A'_n and B'_n self-adjoint – and necessarily commuting, because S_n is normal – and note that $\|A_n - A'_n\| \to 0$ and $\|B_n - B'_n\| \to 0$ because $\{A_n - A'_n\}$ and $\{B_n - B'_n\}$ both belong to $\mathfrak{I}$. However, we do not know if such a normal lift S exits.

To obtain the contradiction we need less: It suffices to find a normal operator T' in $\mathfrak{B}$ within distance $\varepsilon/2$ to $\pi(T)$ such that T' lifts to a normal operator in $\mathfrak{A}$. This is shown in the three propositions below, as we remark that $\mathfrak{B}$ is of real rank zero, stable rank one, and has connected unitary group (these facts are easily seen to hold for matrix algebras, and hence also for $\mathfrak{B}$).

For each $\varepsilon > 0$ let Γ_ε be the one-dimensional grid in the complex plane consisting of those points $x + iy$ where either x or y belongs to $\varepsilon\mathbb{R}$.

Proposition 4.2. *Let T be a normal operator in a unital C^*-algebra $\mathfrak{B}$ of stable rank one. Then for each $\varepsilon > 0$ there is a normal operator $T' \in \mathfrak{B}$ such that* $\mathrm{sp}(T') \subseteq \Gamma_\varepsilon$ *and $\|T - T'\| < \varepsilon$.*

Outline of proof: By the definition of stable rank one, every element in $\mathfrak{B}$ can be approximated by invertible elements in $\mathfrak{B}$. It was shown in [25] that this implies that any *normal* operator can be approximated by *normal* invertible operators. By translation, one obtains that any normal operator can be approximated by normal operators that do not have a given complex number in its spectrum; and hence – by iteration – by normal operators whose spectrum do not intersect any given finite set. Choosing a suitable finite set of points in the holes of the grid Γ_ε one obtains a normal operator S close to T for which there is a continuous function $f\colon \mathrm{sp}(S) \to \Gamma_\varepsilon$ (in fact, a retract), such that $|f(t) - t|$ is small for all t; and we can then take T' to be $f(S)$.

The proposition below was first proved by Lin in [21]; a more direct proof is given in [14].

Proposition 4.3. *Let $\mathfrak{B}$ be a unital C^*-algebra of real rank zero and with connected unitary group. Let $\varepsilon > 0$ be given and let T be a normal operator in $\mathfrak{B}$ with $\mathrm{sp}(T) \subseteq \Gamma_\varepsilon$. Then there is a normal operator T' in $\mathfrak{B}$ with $\mathrm{sp}(T')$ finite such that $\|T - T'\| < \varepsilon$.*

By definition, a C^*-algebra is of real rank zero if any normal element with spectrum contained in the real line (a self-adjoint operator) can be approximated by a normal element with finite spectrum. Passing from the spectrum being a subset of the real line (a self-adjoint operator) to a more general one-dimensional spectrum permitting loops (in our case: a closed subset of Γ_ε), introduces extra complications that, besides making the proof harder, also force us to require that the unitary group be connected.

Proposition 4.4. *Let $\mathfrak{A}$ and $\mathfrak{B}$ be C^*-algebras and let $\pi\colon \mathfrak{A} \to \mathfrak{B}$ be a surjective *-homomorphism. Each normal operator in $\mathfrak{B}$ of finite spectrum lifts to a normal operator in $\mathfrak{A}$.*

Proof. Let T be a normal operator in $\mathfrak{B}$ with finite spectrum, and find continuous functions $f\colon \mathrm{sp}(T) \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{C}$ such that $(g \circ f)(t) = t$ for all $t \in \mathrm{sp}(T)$. Lift the self-adjoint operator $f(T)$ to a self-adjoint operator A in $\mathfrak{A}$; then $g(A)$ is a normal operator in $\mathfrak{A}$ that lifts T. (Note that $g(A)$ not necessarily has finite spectrum.) □

5. High dimensional simple C^*-algebras

In Section 2 we discussed properties of a simple C^*-algebra A, including the somewhat technical notions of almost unperforation of the semigroup of equivalence classes of projections, $V(A)$, and of the Cuntz semigroup, $W(A)$. Until the mid 1990s it was believed that all simple C^*-algebras might enjoy these properties; then Jesper Villadsen constructed a counterexample, [32], by taking an inductive limit of algebras of the form $M_{k(n)}(X^{d(n)})$ for a suitable space X (eg. $X = S^2$) and for suitable increasing sequences $k(n)$ and $d(n)$ of natural numbers. It is a crucial point in the construction

that the connecting mappings $M_{k(n)}(X^{d(n)}) \to M_{k(n+1)}(X^{d(n+1)})$ be chosen in such a way that the inductive limit C^*-algebra becomes simple and – at the same time – that certain high-dimensional properties of the spaces $X^{d(n)}$ are preserved. These techniques of Villadsen have since then been used by several people, including Villadsen himself, to construct many other simple C^*-algebras with various exotic properties, including the example by the author of a simple C^*-algebra with a finite and an infinite projection (and hence a simple stably infinite C^*-algebra which is not purely infinite) as well as various counterexamples to Elliott's classification conjecture (as formulated in Section 3).

5.1. The C^*-algebra associated with a multiplier endomorphism. The construction presented here is a special case of Pimsner's construction of a class of C^*-algebras, called the Pimsner algebras, associated with Hilbert bimodules over C^*-algebras. The construction is implicitly contained in our paper [28], where the reader can find more details. Recall that the multiplier algebra, $\mathcal{M}(A)$, of a C^*-algebra A is the largest unital C^*-algebra that contains A as an essential closed two-sided ideal.

To each pair (A, ρ), where A is a (stable) C^*-algebra and $\rho\colon A \to \mathcal{M}(A)$ is a non-degenerate[1] injective *-homomorphism, we associate a C^*-algebra $C^*(A, \rho)$, which in spirit is the crossed product of A by ρ. (We also use the term *multiplier endomorphism* to denote a *-homomorphism from a C^*-algebra into its multiplier algebra.)

The C^*-algebra $C^*(A, \rho)$ is formally constructed as follows. Since ρ is non-degenerate it extends (uniquely) to a strictly continuous unital *-homomorphism $\rho\colon \mathcal{M}(A) \to \mathcal{M}(A)$. Put

$$B = C^*(A, \rho(A), \rho^2(A), \rho^3(A), \dots) \subseteq \mathcal{M}(A),$$

note that ρ restricts to an endomorphism on B; form the inductive limit

$$B \xrightarrow{\rho} B \xrightarrow{\rho} B \xrightarrow{\rho} \cdots \longrightarrow \overline{B},$$

and extend ρ to an automorphism $\bar{\rho}$ on $\overline{B}$. More explicitly, if $\mu_n\colon B \to \overline{B}$ is the inductive limit map from the nth copy of B, $n \geq 0$, then $\bar{\rho}(\mu_n(b)) = \mu_n(\rho(b)) = \mu_{n-1}(b)$, for $b \in B$. The inverse of $\bar{\rho}$ is given by $(\bar{\rho})^{-1}(\mu_n(b)) = \mu_{n+1}(b)$ for $b \in B$. Put $A_{k-\ell} = \mu_\ell(\rho^k(A))$. Then $A_0 = A$, $\bar{\rho}(A_n) = A_{n+1}$ for all $n \in \mathbb{Z}$, and

$$\overline{B} = C^*(\dots, A_{-2}, A_{-1}, A_0, A_1, A_2, \dots).$$

Let $C^*(A, \rho)$ be the crossed product $\overline{B} \rtimes_{\bar{\rho}} \mathbb{Z}$.

Let u be the canonical unitary in the multiplier algebra of $C^*(A, \rho)$ that implements $\bar{\rho}$. Then $C^*(A, \rho)$ is the closure of the span of elements of the form $a_k u^k$, where

[1] By non-degenerate we mean that ρ maps an approximate unit for A into a sequence that converges strictly to 1 in $\mathcal{M}(A)$.

$k \in \mathbb{Z}$ and $a_k \in C^*(A_n \mid n \in \mathbb{Z})$. $C^*(A_n, A_{n+1}, \ldots, A_m)$ is a closed two-sided ideal in $C^*(A_n, A_{n+1}, \ldots, A_k)$ whenever $n \leq m \leq k$. In particular, $A_0 = A$ is a closed two-sided ideal in $C^*(A_0, A_1)$. If we view $C^*(A_0, A_1)$ as being a sub-C^*-algebra of $\mathcal{M}(A)$, then $\rho(a) = uau^*$ holds for all a in A.

Proposition 5.1. *If A is nuclear, then so is $C^*(A, \rho)$.*

Proof. This follows from the construction of $C^*(A, \rho)$ and the fact that the class of nuclear C^*-algebras is closed under extensions, inductive limits, and crossed products by $\mathbb{Z}$. □

Proposition 5.2. *$C^*(A, \rho)$ is simple if ρ is minimal[2] and if ρ^n is properly outer[3] for every natural number n.*

Proof. If ρ is minimal, then so is the automorphism $\bar{\rho}$ on $\bar{B}$; and if all powers of ρ are properly outer, then the same holds for all powers of $\bar{\rho}$. The proposition therefore follows from [22, Theorem 7.2]. □

If ρ is minimal, and if $\rho(A) \subsetneq A$, as will be the case in the situation we consider in the next subsection, then ρ^n is automatically properly outer for all $n \neq 0$.

We give below conditions that will ensure that a projection p in A is finite, respectively, infinite in $C^*(A, \rho)$.

Proposition 5.3. *Let p be a projection in A.*

(i) *If there exists a projection q in A which is equivalent to p (relatively to A) and is a proper subprojection of $\rho(p)$, then p is infinite in $C^*(A, \rho)$.*

(ii) *If $C^*(A, \rho)$ is simple, and if there is a projection e in A such that $e \not\lesssim p \oplus \rho(p) \oplus \rho^2(p) \oplus \cdots \oplus \rho^n(p)$ (inside $\mathcal{M}(A)$) for all natural numbers n, then p is finite in $C^*(A, \rho)$.*

Recall that we have a canonical unitary u in the multiplier algebra of $C^*(A, \rho)$ that implements ρ, ie., $\rho(a) = uau^*$ for $a \in A$. In particular, $\rho(p) \sim p$ in $C^*(A, \rho)$, so the assumption in (i) implies that p is equivalent to a proper subprojection of itself, and hence that p is infinite. Part (ii) is a more technical and is verified (in a slightly different setting) in [28, Lemma 6.4].

5.2. A simple C^*-algebra with a finite and an infinite projection. We apply the crossed product construction from the previous section to the stable C^*-algebra $A = C(Z) \otimes \mathcal{K} = C(Z, \mathcal{K})$ where Z is the infinite Cartesian product of 2-spheres, $Z = \prod_{n=1}^{\infty} S^2$, and where $\mathcal{K}$ denotes the compact operators on a separable Hilbert space. The multiplier algebra $\mathcal{M}(A)$ coincides in this case with the set of all bounded

[2] ρ is minimal if there are no non-trivial ρ-invariant closed two-sided ideals in A; and an ideal I in A is said to be ρ-invariant if $A\rho(I)A \subseteq I$.

[3] An endomorphism $\rho\colon A \to \mathcal{M}(A)$ is properly outer if its restriction to each non-zero ρ-invariant ideal has norm distance 2 to a multiplier inner endomorphism.

*-strongly continuous functions from Z into $B(H)$, the bounded operators on the Hilbert space on which the compact operators $\mathcal{K}$ acts.

The multiplier endomorphism $\rho\colon A \to \mathcal{M}(A)$ of our construction is of the form $\sum_{j=-\infty}^{\infty} \rho_j$, where each ρ_j is an endomorphism on A, and where the sum $\sum_{j=-\infty}^{\infty} \rho_j(a)$ is strictly convergent for each $a \in A$. (We ensure non-degeneracy of ρ by replacing it with $V^*\rho(\cdot)V$ for some isometry V in $\mathcal{M}(A)$ if necessary.) Each endomorphism ρ_j is induced by a continuous function $Z \to Z$ of the form $(z_1, z_2, \dots) \mapsto (c_1, \dots, c_k, z_{\nu(k+1)}, z_{\nu(k+2)}, \dots)$ (or of the form $(z_1, z_2, \dots) \mapsto (z_{\nu(1)}, z_{\nu(2)}, \dots)$) for suitable $k \in \mathbb{N}$, points $c_i \in S^2$, and for a suitable "shuffle-map" $\nu\colon \mathbb{N} \to \mathbb{N}$ (that all depend on j). The points c_i are chosen such that ρ becomes minimal. As $\rho^n(A) \subsetneq A$ for all n, it follows from Proposition 5.2 that $C^*(A, \rho)$ is simple. The shuffle maps ν (one for each j) are chosen in such a way that certain projections (defined below) have non-trivial Euler class.

If e is a constant 1-dimensional projection in A, then $\rho(e)$ is infinite dimensional and constant, so e is equivalent to a proper subprojection of $\rho(e)$ thus making e infinite in $C^*(A, \rho)$, cf. Proposition 5.3 (i).

It requires more effort to get a finite projection in $C^*(A, \rho)$. For every non-zero projection p in A the projection $\rho(p)$ is a pointwise infinite dimensional in $\mathcal{M}(A)$ (when viewed as a *-strongly continuous function $Z \to B(H)$). We want this projection to be finite in $C^*(A, \rho)$; even more, p must satisfy the condition in Proposition 5.3 (ii) wrt. some projection e.

To this end we take a one-dimensional projection p in $C(S^2, M_2)$ with non-trivial Euler class (p could be the "Bott projection" over S^2). For each $j \in \mathbb{N}$, define $p_j \in C(Z, M_2) \subset A$ by $p_j(z) = p(z_j)$, where $z = (z_1, z_2, \dots) \in Z$; so that p_j is the Bott projection over the jth copy of S^2. For each finite set $I = \{j_1, j_2, \dots, j_k\} \subseteq \mathbb{N}$, let $p_I \in C(Z, M_2 \otimes M_2 \otimes \cdots \otimes M_2) \subseteq A$ be the projection given by

$$p_I(z) = p_{j_1}(z) \otimes p_{j_2}(z) \otimes \cdots \otimes p_{j_k}(z), \quad z \in Z.$$

It is shown in [28] that p_1, the Bott projection over the first copy of S^2 in Z, is a finite projection in $C^*(A, \rho)$. The proof uses the precise definition of the multiplier endomorphism $\rho\colon A \to \mathcal{M}(A)$, Proposition 5.3 (ii) applied to p_1 and with e being a constant one-dimensional projection, and the proposition below (cf. [28, Proposition 3.2]). (Note that if q is a projection in $C(Z, \mathcal{K})$ with non-trivial Euler class then $e \not\precsim q$ by a fundamental property of the Euler class.)

Proposition 5.4. *Let $I_1, I_2, \dots, I_m$ be non-empty finite subsets of $\mathbb{N}$. Then the following conditions are equivalent:*

(i) *The Euler class of $p_{I_1} \oplus p_{I_2} \oplus \cdots \oplus p_{I_m}$ is non-trivial.*

(ii) *For all subsets F of $\{1, 2, \dots, m\}$ we have $\left|\bigcup_{j\in F} I_j\right| \geq |F|$.*

(iii) *There is a matching $t_1 \in I_1, t_2 \in I_2, \dots, t_m \in I_m$.*

Putting these results together we obtain the following main result from [28]:

Theorem 5.5. *The C^*-algebra $C^*(A, \rho)$, with $A = C(Z, \mathcal{K})$, with $Z = \prod_{n=1}^{\infty} S^2$, and with $\rho\colon A \to \mathcal{M}(A)$ being the multiplier endomorphism described above, is simple, separable, nuclear, and it contains an infinite projection and a non-zero finite projection.*

Corollary 5.6. *There is a simple, separable, nuclear C^*-algebra that is stably infinite but not purely infinite; and there is a simple, separable, nuclear, unital, finite C^*-algebra that is not stably finite.*

Proof. The C^*-algebra $B = C^*(A, \rho)$ from Theorem 5.5 is stably infinite (containing an infinite projection) and not purely infinite (because it contains a non-zero finite projection). If p is a non-zero finite projection in B, then pBp is finite but not stably finite. □

5.3. Applications and other examples. The example of a simple C^*-algebra with a finite and an infinite projection as well as other examples constructed later by A. Toms give counterexamples to Elliott's conjecture, or at least they show that the Elliott invariant as defined in Section 3 does not suffice to classify separable nuclear simple (unital) C^*-algebras.

Recall from Section 3 that if A is a stably infinite, simple, unital C^*-algebra, then its Elliott invariant reduces to the triple $(K_0(A), [1_A], K_1(A))$.

Theorem 5.7. *There are simple, separable, nuclear, stably infinite unital C^*-algebras A and B such that*

$$(K_0(A), [1_A], K_1(A)) \cong (K_0(B), [1_B], K_1(B)) \quad \textit{and} \quad A \ncong B.$$

Proof. Let A be as in the first part of Corollary 5.6. There is a purely infinite simple nuclear unital C^*-algebra B such that $(K_0(A), [1_A], K_1(A))$ is isomorphic to $(K_0(B), [1_B], K_1(B))$ (see [27, Proposition 4.3.3 and 4.3.4]). As B is purely infinite and A is not, the two C^*-algebras are not isomorphic. □

Note also that it follows from Proposition 2.7 that the C^*-algebra $C^*(A, \rho)$ from Theorem 5.5 is tensorially prime (see Subsection 2.3).

Toms used Villadsen's techniques to construct simple stably finite (AH- and ASH-algebras) with explicit strong perforation in K_0 (eg. with (K_0, K_0^+) isomorphic to $(\mathbb{Z}, S)$ where S can be almost any subsemigroup of $\mathbb{Z}^+$ with $S - S = \mathbb{Z}$). Recently, Toms also constructed ingenious counterexamples to Elliott's conjecture in the stably finite case, ie. pairs of non-isomorphic simple, separable, nuclear, stably finite C^*-algebras with the same Elliott invariant (and for this matter also other invariants, not normally included in the Elliott invariant) (see [31]).

References

[1] Avitzour, D., Free products of C^*-algebras. *Trans. Amer. Math. Soc.* **271** (1982), 423–435.

[2] Blackadar, B., and Cuntz, J., The structure of stable algebraically simple C^*-algebras. *Amer. J. Math.* **104** (1982), 813–822.

[3] Blackadar, B., and Handelman, D., Dimension functions and traces on C^*-algebras. *J. Funct. Anal.* **45** (1982), 297–340.

[4] Blackadar, B., and Rørdam, M., Extending states on preordered semigroups and the existence of quasitraces on C^*-algebras. *J. Algebra* **152** (1992), 240–247.

[5] Brown, L. G., and Pedersen, G. K., C^*-algebras of real rank zero. *J. Funct. Anal.* **99** (1991), 131–149.

[6] Cuntz, J., K-theory for certain C^*-algebras. *Ann. of Math.* **113** (1981), 181–197.

[7] Dădărlat, M., and Gong, G., A classification result for approximately homogeneous C^*-algebras of real rank zero. *Geom. Funct. Anal.* **7** (1997), 646–711.

[8] Dykema, K. J., Exactness of reduced amalgamated free product C^*-algebras. *Forum Math.* **16** (2) (2004), 161–180.

[9] Dykema, K. J., and Rørdam, M., Projections in free product C^*-algebras. *Geom. Funct. Anal.* **8** (1998), 1–16.

[10] —, Purely infinite simple C^*-algebras arising from free product constructions. *Canad. J. Math.* **50** (2) (1998), 323–341.

[11] Dykema, K. J., and Shlyakhtenko, D., Exactness of Cuntz-Pimsner C^*-algebras. *Proc. Edinburgh Math. Soc.* (2) **44** (2) (2001), 425–444.

[12] Elliott, G. A., On the classification of C^*-algebras of real rank zero. *J. Reine Angew. Math.* **443** (1993), 179–219.

[13] Elliott, G. A., Gong, G., and Li, L., On the classification of simple inductive limit C^*-algebras II: The isomorphism theorem. Preprint, 1998.

[14] Friis, P., and Rørdam, M., Almost commuting self-adjoint matrices — A short proof of Huaxin Lin's theorem. *J. Reine Angew. Math.* **479** (1996), 121–131.

[15] Goodearl, K. R., and Handelman, D., Rank functions and K_0 of regular rings. *J. Pure Appl. Algebra* **7** (1976), 195–216.

[16] Haagerup, U., Every quasi-trace on an exact C^*-algebra is a trace. Preprint, 1991.

[17] Haagerup, U., and Thorbjørnsen, S., Random matrices and K-theory for exact C^*-algebras. *Documenta Math.* **4** (1999), 341–450.

[18] Jiang, X., and Su, H., On a simple unital projectionless C^*-algebra. *Amer. J. Math.* **121** (2) (1999), 359–413.

[19] Kirchberg, E., The classification of purely infinite C^*-algebras using Kasparov's Theory. In preparation.

[20] Lin, H., Almost commuting self-adjoint matrices and applications. In *Operator Algebras and their Applications* (ed. by P. A. Fillmore and J. A. Mingo), Fields Inst. Commun. 13, Amer. Math. Soc., Providence, RI , 1995, 193–233.

[21] —, Approximation by normal elements with finite spectra in C^*-algebras of real rank zero. *Pacific J. Math.* **173** (2) (1996), 443–489.

[22] Olesen, D., and Pedersen, G. K., Applications of the Connes Spectrum to C^*-dynamical Systems, III. *J. Funct. Anal.* **45** (3) (1981), 357–390.

[23] Phillips, N. C., A Classification Theorem for Nuclear Purely Infinite Simple C^*-Algebras. *Documenta Math.* **5** (2000), 49–114.

[24] Rieffel, M., Dimension and stable rank in the K-theory of C^*-algebras. *Proc. London Math. Soc.* **46** (3) (1983), 301–333.

[25] Rørdam, M., Advances in the theory of unitary rank and regular approximation. *Ann. of Math.* **128** (1988), 153–172.

[26] —, Stability of C^*-algebras is not a stable property. *Documenta Math.* **2** (1997), 375–386.

[27] —, Classification of Nuclear, Simple C^*-algebras. In *Classification of Nuclear C^*-Algebras. Entropy in Operator Algebras* (ed. by J. Cuntz and V. Jones), Encyclopaedia Math. Sci. 126, Operator Algebras and Non-commutative Geometry 7, Springer-Verlag, Berlin 2001, 1–145.

[28] —, A simple C^*-algebra with a finite and an infinite projection. *Acta Math.* **191** (2003), 109–142.

[29] —, The stable and the real rank of $\mathcal{Z}$-absorbing C*-algebras. *Internat. J. Math.* **15** (10) (2004), 1065–1084.

[30] Rosenberg, J., and Schochet, C., The Künneth Theorem and the Universal Coefficient Theorem for Kasparov's generalized K-functor. *Duke Math. J.* **55** (2) (1987), 431–474.

[31] Toms, A., On the classification problem for nuclear C^*-algebras. Preprint; math. archive math.OA/0509103.

[32] Villadsen, J., Simple C^*-algebras with perforation. *J. Funct. Anal.* **154** (1) (1998), 110–116.

[33] —, On the stable rank of simple C^*-algebras. *J. Amer. Math.* Soc. **12** (4) (1999), 1091–1102.

[34] Winter, W., On the classification of simple $\mathcal{Z}$-stable C^*-algebras with real rank zero and finite decomposition rank. Preprint; math. archive math.OA/0502181.

[35] Zhang, S., A property of purely infinite simple C^*-algebras. *Proc. Amer. Math. Soc.* **109** (1990), 717–720.

Department of Mathematics and Computer Science, University of Southern Denmark,
Campusvej 55, 5230 Odense M, Denmark
E-mail: mikael@imada.sdu.dk

Convexity, complexity, and high dimensions

Stanislaw J. Szarek*

Abstract. We discuss metric, algorithmic and geometric issues related to broadly understood complexity of high dimensional convex sets. The specific topics we bring up include metric entropy and its duality, derandomization of constructions of normed spaces or of convex bodies, and different fundamental questions related to geometric diversity of such bodies, as measured by various isomorphic (as opposed to isometric) invariants.

Mathematics Subject Classification (2000). Primary 46B20; Secondary 46B09, 47B06, 52A21, 52C17, 15A52, 90C25, 94B75.

Keywords. Convex body, high dimension, complexity, metric entropy, asymptotic geometric analysis.

1. Introduction

When modeling complex (real-life or abstract) systems with many degrees of freedom, we are frequently led to mathematical objects whose dimension can be related to the number of free parameters in the underlying system and, as a consequence, is very large. Since many naturally appearing relationships between, or constraints on the parameters are linear or at least convex, we are thus led to high dimensional convex sets. Two areas of traditional mathematics that come to mind when faced with the problem of analyzing such sets are *classical geometry* and *functional analysis*. However, geometry is usually focused on obtaining very precise information for a fixed, not-too-large dimension. Functional analysis, on the other hand, is typically concerned with the infinite-dimensional setting (which frequently is an idealization of a very large dimension), but often provides only qualitative information. Fortunately, there is a middle ground between these two approaches and it has turned out to be quite fertile. The last few decades witnessed the development of a quite powerful quantitative methodology in geometric functional analysis that, together with similar advances in areas such as combinatorics or theoretical computer science, has been lately referred to as *asymptotic geometric analysis*. In a nutshell, the prescription for success of the asymptotic theory depends on identifying and exploiting *approximate* symmetries of various problems that escaped the earlier "too qualitative" or "too rigid" methods of classical functional analysis and classical geometry. More specifically,

*Supported in part by a grant from the National Science Foundation (U.S.A.).

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

an important feature of the area is the predominantly *isomorphic* (as opposed to *isometric*) character of the questions: one is usually after the rough (i.e., "up to a universal constant") asymptotic order of the quantities studied and not their more precise behavior. This sometimes makes the problem solvable. [See the article [48] by B. Klartag in this collection for a discussion of developments in the so-called "almost isometric" theory.] However, universal estimates are required, independent of the particular instance of the problem, most notably of the dimension. As is very well known to specialists, but perhaps not fully appreciated by non-experts, this last feature is absolutely crucial because it allows for applications to infinite dimensional functional analysis and to quantitative questions in applied fields. This framework led to discoveries of many surprising phenomena, which may be subsumed in the following "experimental" observation: low-dimensional intuitions are often *very* wrong in high dimensions. However, as opposed to other fields such as topology, high-dimensional curiosities generally do not appear via a quantum jump; say, when passing from dimension 3 to 4, or from 7 to 8. Instead, small changes accumulate as the dimension increases, leading ultimately to a qualitatively new picture.

In the present article we shall focus on those aspects of the theory that are relevant to broadly defined *complexity* of convex sets. To exemplify what we mean by complexity, we will now hint at three alternative viewpoints on the notion. Below K will stand for a generic *convex body* in $\mathbb{R}^n$ (i.e., compact convex set with nonempty interior).

(i) *The algorithmic complexity.* How difficult is it to describe K? A prime example (which we mention mainly for demonstration purposes) is here the algorithmic complexity of the *membership oracle*: How difficult is it to decide whether a point belongs to K? Another class of questions is: Suppose that the existence of K with certain properties is given by a non-constructive proof, or by probabilistic considerations; is it possible to give an explicit example, or an efficient derandomized algorithm?

(ii) *The geometric complexity, or diversity.* How complicated (in an appropriate geometric sense) is K? To what extent do convex bodies of the same dimension exhibit common features? In particular, to what extent do they resemble the arguably most regular body, the Euclidean ball?

(iii) *The metric entropy.* Here we need an underlying metric structure, typically given by a norm. Given $\varepsilon > 0$, how many balls of radius ε do we need to cover K? The logarithm (to base 2) of the minimal cardinality of such cover is called the *metric entropy function* of K and can be interpreted as the complexity of K, measured in bits, at the level of resolution ε with respect to the metric in question.

In what follows we will provide some background information concerning these three aspects of complexity, describe recent developments in each area, and list related open problems. Each of the viewpoints (i)–(iii) will correspond to one section in the exposition; however, we will reverse the order. On the other hand, we emphasize that the three points of view are intimately interconnected. For example, one aspect of geometric diversity involves approximating convex bodies by simple ones, a problem which has obvious algorithmic ramifications. A sample such question, approximating by polytopes, has been extensively studied and reported on in [14], see also [92].

Notation will be introduced as we proceed, but we list below a few general rules and conventions that will be used throughout the paper. We will sometimes use unexplained (but standard in the field) notation in side remarks. For this and for more background on the issues discussed here we refer the reader to the monographs [75], [108], [84] and, for more up-to-date expositions, to surveys contained in [41], especially the chapters [42], [18], [24], [30], [43], [50], [56], [59], [63], [64], and – particularly for the motivational aspects – to the 1996 ECM and the 1998 ICM talks by V. Milman [72], [73].

If X is a normed space, we will write B_X for its unit ball (centered at the origin). In the other direction, if K is a convex body in $\mathbb{R}^n$ containing the origin in its interior, $\|\cdot\|_K$ will stand for the gauge of K (i.e., $\|x\|_K := \inf\{t > 0 : x \in tK\}$); if K is symmetric with respect to the origin, the gauge of K is just the norm for which K is the unit ball. Thus, in a way, the body K (symmetric with respect to the origin) is identified with the normed space $(\mathbb{R}^n, \|\cdot\|_K)$; this is the main reason for the interplay of geometric and functional analytic ideas.

The letters $C, c_0, C', \ldots$ will stand for absolute positive constants, independent of the particular instance of the problem considered, most notably of the dimension. However, the numerical values corresponding to the same symbol may vary between occurrences. Similarly, $C(\alpha)$ will denote a constant depending only on the parameter α, and so on. For two functions f, g (depending on the same or on different parameters), $f \sim g$ will mean that f and g are of the same order, i.e., $cf \le g \le Cf$ (with $C, c > 0$ independent of the parameters involved, as required by the previous convention).

2. Metric entropy and its duality

While in many applications *calculating* the metric entropy of *specific* sets is the primary objective, here we will mention applications only in passing and concentrate instead on the more fundamental properties of the notion, particularly those connected to duality considerations.

2.1. Notation and historical background, the duality conjecture. If K, B are subsets of a vector space, the *covering number* of K by B, denoted $N(K, B)$, is the minimal number of translates of B needed to cover K. Similarly, the *packing* number $M(K, B)$ is the maximal number of disjoint translates of B by elements of K. [Note that geometers usually require only that the *interiors* of the translates be disjoint.] The two concepts are closely related, particularly if B is centrally symmetric; we have then $N(K, 2B) \le M(K, B) \le N(K, B)$. If B is a ball in a normed space and K a subset of that space (the setting and the point of view we will usually employ), these notions reduce to considerations involving the smallest ε-nets or the largest ε-separated (or 2ε-separated) subsets of K.

Besides the immediate geometric framework, packing and covering numbers appear naturally in numerous subfields of mathematics, ranging from classical and functional analysis through operator theory and probability theory (particularly when studying stochastic processes, see [26], [95], [61], [53]) to information theory and computer science (where, for example, a code is typically a packing). As with other notions touching on convexity, an important role is played by considerations involving duality. The central problem in this area is the 1972 *duality conjecture for covering numbers* due to Pietsch [79], which has been originally formulated in the operator-theoretic context (see below), but which in the present setting can be stated as

Conjecture 2.1 (*The Duality Conjecture*). There exist numerical constants $a, b \geq 1$ such that for any dimension n and for any two symmetric convex bodies K, B in $\mathbb{R}^n$ one has

$$b^{-1} \log N(B^\circ, aK^\circ) \leq \log N(K, B) \leq b \log N(B^\circ, a^{-1}K^\circ). \tag{1}$$

Above and in what follows $A^\circ := \{u \in \mathbb{R}^n : \sup_{x\in A} \langle x, u\rangle \leq 1\}$ is the polar body of A; "symmetric" is a shorthand for "symmetric with respect to the origin" and, for definiteness, all logarithms are to the base 2. For simplicity, we will generally restrict our attention to symmetric sets; however, most statements can be formulated without the symmetry assumption. Of course, due to the bipolar theorem, K and B are exchangeable in (1) and so it is enough to prove only one of the two inequalities. We emphasize that the "equivalence" of the quantities in (1), and overall in this section, is more involved than the relation $\sim$ defined in the introduction.

In our preferred setting of a normed space X, the proper generality is achieved by considering $\log N(K, t\,B_X)$, where $t > 0$ and K is a general (convex, symmetric) subset of X. The polars should then be thought of as subsets of X^*, with $(B_X)^\circ$ coinciding with B_{X^*}, the unit ball of that space, and (1) becomes

$$b^{-1} \log N(B_{X^*}, atK^\circ) \leq \log N(K, tB_X) \leq b \log N(B_{X^*}, a^{-1}tK^\circ). \tag{2}$$

With minimal care, infinite-dimensional spaces and sets may be likewise considered. To avoid stating boundedness/compactness hypotheses, which are peripheral to the phenomena in question, it is convenient to allow $N(\cdot,\cdot)$, $M(\cdot,\cdot)$ etc. to take the value $+\infty$. Finally, the original operator-theoretic formulation of the conjecture is as follows. Given (linear bounded, or compact) operator $u\colon Y \to X$ between two normed spaces we define *entropy numbers* of u as $e_k(u) := \inf\{\varepsilon > 0 : N(uB_Y, \varepsilon B_X) \leq 2^{k-1}\}$. Do we have then

$$a^{-1} e_{bk}(u) \leq e_k(u^*) \leq a e_{k/b}(u^*) \tag{3}$$

(where $u^*\colon X^* \to Y^*$ is the adjoint of u) uniformly over spaces X, Y, operators u and $k \geq 1$?

2.2. The Hilbert space case. To indicate where the difficulty of the problem lies, we shall comment on the enlightening special cases when, in the language of (3), X and/or Y is a Hilbert space. If *both* X and Y are Hilbert spaces, the situation is nearly trivial. Indeed, entropy numbers of a Hilbert space operator depend only on its singular numbers, and the singular numbers of an operator and its adjoint are identical. In the setting of (1), this corresponds to the bodies K, B (and hence K°, B°) being ellipsoids with the pair (K, B) affinely equivalent to the pair (B°, K°) (in that order!). As a consequence, (1), (3), and the appropriate version of (2), hold with $a = b = 1$. At the other extreme, in the general case, the four bodies K, B, K°, B° appearing in (1) may be *all* very different and so the reasons for the duality result (if it indeed does hold) must be much deeper. Finally, if one of the spaces (say, X) is a Hilbert space, looking at the equivalence (2) we see that it expresses what seems to be a rather fundamental property of *all* convex subsets of the Hilbert space.

Let us also point out that if $\dim X = \dim K := k < \infty$ (and K is bounded), then standard considerations show that, as $t \to 0^+$, both metric entropy functionals $\log N(K, t\, B_X)$ and $\log N(B_{X^*}, t\, K^\circ)$ are equivalent to $k \log(1/t)$. In fact, it is even true that the differences between the functionals and $k \log(1/t)$ are bounded. However, the bounds for the two differences depend in an intricate way on K, not allowing for any meaningful quantitative inferences, nor for deriving any conclusions about reasonably general infinite dimensional sets. [Comments in this paragraph are not dependent on X being a Hilbert (i.e., Euclidean) space.]

2.3. Duality results. The three decades following the statement of the conjecture brought many useful partial and/or related results, see [91], [51], [107], [19], [78], [33], [85] and their references. However, only in the last few years substantial progress was achieved with respect to the original problem. We have (see [8], [9])

Theorem 2.2. *There exist universal constants $a, b \geq 1$ such that* (2) *holds if X is a Hilbert space, uniformly over all symmetric convex sets $K \subset X$ and over $t > 0$. Moreover, the same is true if X is K-convex, with a, b depending on the K-convexity constant of X.*

The notion of *K-convexity* (the notation which, by the way, has nothing to do with our convex set K) goes back to [65] and is well known to specialists; we refer to [84], [64] for a precise definition, background and properties. Requiring K-convexity imposes a rather mild geometric restriction on the underlying space. For example, the class of K-convex spaces includes all L_p-spaces for $1 < p < \infty$ (classical or non-commutative), and similarly all uniformly convex and all uniformly smooth spaces. While many interesting descriptions of this class are possible (see [56], [64]), here we just mention that K-convexity is equivalent (see [82]) to the absence of large subspaces resembling (in the sense of the next section) finite-dimensional ℓ_1-spaces, and that it can be nicely quantified: there is a parameter called the *K-convexity constant*, which can be defined both for finite and infinite dimensional normed spaces, and which has good permanence properties with respect to standard functors of functional analysis.

Translating Theorem 2.2 to the other formulations is straightforward. For example, we get that (3) holds if one of the spaces X, Y is a Hilbert space or, more generally, is K-convex. Similarly, if (say) B is an ellipsoid, then (1) holds uniformly over $n \in \mathbb{N}$ and over (symmetric convex bodies) $K \subset \mathbb{R}^n$ etc. Regarding constants, we know how to prove (2) for the Hilbert space with any $a > 2$, with $b = b(a)$; improvements (to "any $a > 1$") would follow if a certain geometric statement conjectured in [76] was true.

2.4. The convexified packing. An interesting feature of [9] is the formal introduction and an initial study of a modified notion of packing that has already been implicit in [19]. We will provide now some details since this will allow us to present a sketch of the proof of Theorem 2.2 and to pinpoint the ingredients that are missing in the general case of Conjecture 2.1; at the same time, the new notion appears to be interesting by itself. A sequence $x_1, \ldots, x_m$ is called a *convexified B-packing* iff

$$(x_j + B) \cap \operatorname{conv} \bigcup_{i<j} (x_i + B) = \emptyset$$

for $j = 2, \ldots, m$. [We emphasize that, as opposed to the usual notions of packing and covering, the *order* of the points *is* important here.] Next, the *convexified packing number* $\hat{M}(K, B)$ is the maximal length of a sequence in K which is a convexified B-packing. It turns out that for this modified notion the duality in the sense analogous to (1)–(3) does hold, and that it is (essentially) a consequence of a Hahn–Banach type separation theorem. For example, we have (see [9])

If $K, B \subset \mathbb{R}^n$ are convex symmetric bodies, then $\hat{M}(K, B) \le \hat{M}(B^\circ, K^\circ/2)^2$.

Accordingly, if we knew that the numbers $M(\cdot, \cdot)$ and $\hat{M}(\cdot, \cdot)$ were in the appropriate sense equivalent, the original duality conjecture would follow immediately. While clearly $\hat{M}(K, B) \le M(K, B)$, any general inequality going in the opposite direction appears at the first sight unlikely. However, the following reduction (shown in [8]) simplifies the matter substantially.

For a given space X, if (2) *holds for $t = 1$ and all $K \subset X$ verifying $B_X/4 \subset K \subset 4B_X$, then it holds* (*perhaps with different $a, b > 0$*) *for all K and all $t > 0$.*

In other words, it is enough to prove (1) (or even the first inequality in (1)) for K, B such that $B/4 \subset K \subset 4B$. This reduction allows us to close the loop, at least under some additional mild geometric assumptions about the ambient normed space.

For bounded sets in a K-convex space X, $M(\cdot, B_X)$ and $\hat{M}(\cdot, B_X)$ are comparable.

More precisely, we have $\log M(T, B_X) \le \beta \log \hat{M}(T, B_X/4)$ where β depends only on the diameter of the convex set T and (the upper bound on) the K-convexity constant of X. Moreover, the roles of B and T can be reversed if T is symmetric and $T \supset rB$

for some $r > 0$ (with β depending on r). Proofs of both these facts (contained in [9]) are based on the so-called Maurey's lemma (see [81]) and on the ideas from [19].

Theorem 2.2 follows now by combining (more or less) formally the three statements above. The argument suggests several natural questions.

Problem 2.3. Are the quantities $M(\cdot, B_X)$ and $\hat{M}(\cdot, B_X)$ always (in the appropriate sense) comparable? Comparable uniformly over well-bounded subsets of (an arbitrary) normed space X? Comparable uniformly over subsets of a Hilbert space X (without restriction on the diameter)?

An affirmative answer to the second question would imply an affirmative answer to Conjecture 2.1 in full generality. Similarly, an affirmative answer for a specific non-K-convex space X would imply the form (2) of the conjecture for that space (and the form (3), with the second space Y arbitrary). An interesting test case is $X = \ell_1$.

While equivalence of $M(\cdot, B_X)$ and $\hat{M}(\cdot, B_X)$ over all convex subsets of X (i.e., in absence of a uniform upper bound on the diameter, the first and the third part of Problem 2.3) is not required for the corresponding case of the Duality Conjecture 2.1, good understanding of the relationship between the two quantities may have implications for complexity theory. Indeed, a standard device in constructing geometric algorithms is a *separation oracle* (cf. [35]): if T is a convex set then, for a given x, the oracle either attests that $x \in T$ or returns a functional efficiently separating x from T. It is arguable that quantities of the type $\hat{M}(T, \cdot)$ correctly describe complexities of the set T with respect to many such algorithms.

3. Geometric complexity of convex bodies and their diversity

When comparing shapes of convex bodies, it is most natural in our context to not distinguish K from its images via invertible affine maps. This may be thought of as choosing for each body the coordinate system that is most appropriate for the particular property that is being studied, and leads to the concept of the *Banach–Mazur distance*. For (symmetric) convex bodies $U, V \subset \mathbb{R}^n$ one sets

$$d(U, V) := \inf\{\lambda > 0 : \text{there exists } w \in \mathrm{GL}(n) \text{ such that } U \subset wV \subset \lambda U\}.$$

This definition is usually formulated in the language of normed spaces: $d(X, Y) := \inf\{\|w\| \cdot \|w^{-1}\| : w\colon X \to Y \text{ an isomorphism}\}$. We refer to the monograph [108] for an exhaustive study of issues related to this notion.

3.1. The structure of the Banach–Mazur compactum. The set of (classes of affinely equivalent) symmetric convex bodies in $\mathbb{R}^n$ (or, equivalently, of classes of isometric n-dimensional normed spaces), endowed with the Banach–Mazur distance, is usually called the (nth) *Banach–Mazur compactum* or the *Minkowski compactum*. [See [1] for most recent results about the *topological* structure of this set.] It is actually

$\log d(\cdot,\cdot)$ which has the usual properties of a distance function, but it is customary to abuse the notation and talk about $d(\cdot,\cdot)$ as if it was a metric. It is a fundamental result due to F. John [40]) that for any n-dimensional symmetric convex body K its distance $d(K, B_2^n)$ to the Euclidean ball B_2^n may be at most $\sqrt{n}$, and it is easy to see that this bound can not be in general improved. It follows right away that the *diameter* of the compactum is at most n, and the remarkable result of Gluskin [31] shows that this bound can not be substantially improved: we do have pairs K, B of n-dimensional symmetric convex bodies for which $d(K, B) \geq cn$ (where $c > 0$ is, according to our convention, a universal constant independent of n). We take this opportunity to point out several unsolved problems in this general direction. First, it would be interesting to determine the *exact* diameter of the Banach–Mazur compactum for specific low dimensions; in fact, the only case when the diameter of the compactum is precisely known is $n = 2$ (see [11]), and the complex analogue is unknown even in dimension 2. A more serious problem is the question of finding (the order of) the maximal distance of specific important convex bodies to general ones of the same dimension. For the *n-dimensional cube* B_∞^n (the unit ball of ℓ_∞^n), the easy lower and upper bounds of $\sqrt{n}$ and n were improved only around 1990 in, respectively, [98] and [20]. The lower bound $cn \log n$ from [98] remains the best known, while the upper one has been tightened in [101] and in a series of papers by Giannopoulos culminating in $Cn^{5/6}$ in [29]. Clearly, a wide gap still persists. Another wide open question is that about the diameter of the compactum of the *not-necessarily-symmetric* n-dimensional convex bodies, with the definition of the distance involving additionally a minimum over translations. The analogue of John's result (also contained in [40]) yields the value n for the radius with center at B_2^n. The resulting estimate n^2 on the diameter has been improved in [86] (see also [13]) to $n^{4/3}$ (times a logarithmic factor). For a more general discussion of the isomorphic theory of non-symmetric convex bodies we refer to the recent articles [54], [74], [34] and their references.

3.2. Quotient of a subspace theorem and its aftermath. It follows from the results quoted in Section 3.1 that convex bodies/normed spaces can be quite distant in the Banach–Mazur sense. Accordingly, it was a major surprise when Milman ([69]) discovered in the mid 1980s that, in some sense, every convex body hides somewhere inside its structure an ellipsoid of nearly full dimension. More precisely, we have (in the language of normed spaces)

Theorem 3.1 (Quotient of a subspace theorem). *Given $\theta \in (0, 1)$ and an n-dimensional normed space X there exists a* subspace of a quotient *of X whose dimension is $\geq \theta n$ and whose Banach–Mazur distance to the Euclidean space does not exceed $C(\theta)$. Moreover, $C(\theta)$ can be chosen to verify* $\lim_{\theta\to 0^+} = 1$.

In other words, every n-dimensional symmetric convex body admits a central section and an affine image (not necessarily bijective) of that section which is of dimension $\geq \theta n$ and which is $C(\theta)$-equivalent (in the sense of the Banach–Mazur distance) to a Euclidean ball. Theorem 3.1 should be compared with the much earlier

celebrated Dvoretzky theorem (see [27], and [68] for the improved version quoted here) which, in the same context, asserts existence of almost Euclidean *sections*, or subspaces, whose dimension is just of order $\log n$. It is easy to see that, in general, if we only use the operation of passing to a subspace (or, dually, only the operation of passing to a quotient), then this logarithmic order can not be improved. [We discuss some remarkable special cases when it can be dramatically improved in the next section.] Still, it is conceivable that *some* considerable regularity can be achieved by a single operation of passing to a "proportional" subspace (or to a "proportional" quotient). In the wake of his quotient of a subspace theorem Milman stated in his 1986 ICM lecture ([70]) several specific problems going in that direction. All these problems were recently answered in the negative due to the discovery of a new phenomenon which we will next describe.

3.3. The saturation phenomenon. The following result from [103] is a sample illustration of the phenomenon.

Theorem 3.2 (The saturation phenomenon). *Let n and m_0 be positive integers with $\sqrt{n}\,\log n \le m_0 \le n$. Then, for every finite dimensional normed space W with*

$$\dim W \le c_1 m_0/\sqrt{n} \tag{4}$$

there exists an n-dimensional normed space X such that every subspace Y of X with $\dim Y \ge m_0$ contains a contractively complemented subspace isometric to W.

Loosely speaking, Theorem 3.2 says that the space X is so "saturated" with subspaces isometric to W (copies of W), that such subspaces persist in every "sufficiently large" subspace of X. Furthermore, due to the complementability clause in the assertion of Theorem 3.2, the statement can be dualized, i.e., "every subspace Y of X" can be replaced by "every quotient Y of X." Thus, in general, passing to large subspaces *or* large quotients can not erase k-dimensional features of a space if k is below certain threshold value. This is in stark contrast to the operation of passing to a large quotient of a subspace, which – by Theorem 3.1 – may lead, in a sense, to losing *all* information about the original space, no matter how complicated that space was.

Let us point out that, under the hypotheses of the Theorem, the dimension $k := \dim W$ is always nontrivial (i.e., large, if n is large), and in the most interesting case when $m_0 \sim n$ (say, $m_0 \approx n/2$) we can have $k \sim \sqrt{n}$. It follows that Theorem 3.2 imposes strict limits on properties that may be achieved (or improved) by passing to a large subspace (resp., quotient). Indeed, no property of normed spaces whose violation can be witnessed inside subspaces of dimension $\ll \sqrt{\dim X}$ (and which is inherited by complemented subspaces of a space) can be in general achieved by passing to a subspace (resp., quotient) of X whose dimension is comparable to that of X. To demonstrate that, we choose any space W with $\dim W \ll \sqrt{n}$ which does not have the property in question and use Theorem 3.2 to construct an n-dimensional space; then every sufficiently large subspace (resp., quotient) of X contains a contractively

complemented subspace isometric to W and consequently can not have our property. Examples of properties which can be so "prevented" include being of nontrivial type or cotype, which immediately settles in the negative (and in a very strong sense) Problem 1 from [70]: *Does every n-dimensional normed space admit a quotient of dimension $\geq n/2$ whose cotype 2 constant is bounded by a universal numerical constant?*

Statements similar to Theorem 3.2 hold if it is additionally required that X has certain regularity properties. For example, if we insist that the cotype q constant of X (for some $q \in (2, \infty)$) be controlled, it is possible to construct X which is saturated with copies of any given space W whose cotype q constant is not too large, provided $\dim W$ verifies a condition resembling (4). These topics were developed in [103], [104] and led to negative answers to Problems 2 and 3 from [70].

All the above notwithstanding, some *global* regularity of bodies/spaces may be achievable by passing to proportional quotient or subspaces. For example, already in [69], as a step in the proof of Theorem 3.1, it was established that every finite dimensional normed space admits a "proportional" quotient of well-bounded *volume ratio*, a volumetric characteristic of a convex body closely related to cotype 2 property of the corresponding normed space. It would be important to find more examples of, and/or limitations on such results. As a sample problem we mention the following

Problem 3.3. Given a finite dimensional normed space X, does there exist a subspace $Y \subset X$ with $m := \dim Y \geq \dim X/2$ and a basis $y_1, \ldots, y_m$ of Y such that, denoting by $y_1^*, \ldots, y_m^*$ the dual basis of Y^* we have

$$\mathrm{Ave}_{\varepsilon_i=\pm 1} \left\| \sum_{i=1}^{m} \varepsilon_i y_i \right\| \cdot \mathrm{Ave}_{\eta_j=\pm 1} \left\| \sum_{j=1}^{m} \eta_j y_j^* \right\| \leq Cm \,?$$

It is in fact an open problem whether a similar property holds for every normed space *without* passing to a subspace. [For example, a slightly weaker form of this last question was stated as Problem 6 in [70].]

3.4. Products of convex bodies and the nontrivial projection problem. A measure of geometric complexity of a high dimensional convex body K is whether it can be "reduced," in some meaningful sense, to bodies of substantially lower dimension. One such natural reduction would be approximating K, in the sense of Banach–Mazur distance, by Cartesian products of (two or more) convex bodies, each of which has a reasonably large dimension. The so phrased problem makes sense also for non-symmetric bodies, but in the symmetric case it reduces to the well-known *nontrivial projection problem*.

Problem 3.4. Do there exist $C > 0$ and a sequence $k_n \to +\infty$ such that for every n-dimensional normed space X there is a projection P on X with $\|P\| \leq C$ and $\min\{\mathrm{rank}\, P, \mathrm{rank}(I - P)\} \geq k_n$?

A question about somewhat stronger property, the *finite-dimensional basis problem* was resolved in early 1980s (after having been open for about 50 years) in [32] and [97] (see also [60]), where it was shown that the statement from Problem 3.4 can not hold with k_n substantially larger than $\sqrt{n}$ (more precisely, with $k_n \gg \sqrt{n \log n}$) and that, in general, we can not find projections on X whose rank and corank are of the same order as $\dim X$ and whose norm is $o(\sqrt{\dim X})$. [Note that every k-dimensional subspace of a normed space is complemented via a projection of norm $\leq \sqrt{k}$, see [44], or even slightly smaller, see [52], [50].] Various versions of Problem 3.4 were stated in the ICM talks by Milman (1986, [70]) and, most notably, by Pisier (1983, [83]), with the latter reporting also on the definitive treatment of the case when $\dim X = \infty$: it may then happen that, for any finite rank projection P on X one has $\|P\| \geq c\sqrt{\operatorname{rank} P}$. [The purely infinite dimensional counterexample to splitting into a nontrivial Cartesian product, or even to a weaker property, is provided by the Gowers–Maurey *hereditarily indecomposable* spaces, see [63].]

In spite of all these negative results it is still conceivable that the answer to Problem 3.4 is affirmative, even with $k_n \sim \sqrt{n}$; this threshold is precise (on the power scale) in the case of "the usual suspects," Gluskin-type random spaces, (see [60]) and – perhaps for a reason – parallels some thresholds related to the saturation phenomenon from Section 3.2. In fact, improvements on the extremal order $\sqrt{\operatorname{rank} P}$ for norms of projections have been known for quite a while, see [80], [83]. The following bound ([105]) can be obtained by combining known techniques (in a not-so-straightforward manner, though).

Theorem 3.5. *There exist $C, c > 0$ and a sequence $k_n \geq \exp(c\sqrt{\log n})$ such that, for every n-dimensional normed space X, there is a projection P on X with* $\min\{\operatorname{rank} P, \operatorname{rank}(I - P)\} \geq k_n$ *and* $\|P\| \leq C(\log k_n)^2$.

Going even further, we do not see easy counterexamples to the following (sample) statement stronger than the one in Problem 3.4: *Given an n-dimensional normed space X and an integer m with $\sqrt{n} < m \leq n$, the space X can be split into a direct sum of m subspaces $E_1, \ldots, E_m$ of approximately equal dimensions, and such that if P_j is the projection onto E_j that annihilates all E_i with $i \neq j$, then* $\max_{1 \leq j \leq m} \|P_j\| \leq C$. This would be a generalization of the classical Auerbach lemma which asserts that the answer is yes, with $C = 1$, if $m = n$. However, it is possible that, at least for some range of m, an argument in the spirit of [98] may yield a counterexample.

4. Algorithmic complexity and derandomization, pseudorandom matrices

Many results in asymptotic geometric analysis, including virtually all cited in the preceding section, have been obtained by *probabilistic* considerations. For example, when the objective is to prove the existence of a convex body (or a normed space) with certain property, the strategy is to come up with an appropriate random variable

whose values are convex bodies, and then to show that with nonzero (and typically close to 1) probability the property in question is satisfied. [The arguments usually involve precise metric entropy estimates for various subsets of $\mathbb{R}^n$, or for sets of operators on $\mathbb{R}^n$, combined with large deviation and, particularly, small ball estimates for vector-valued random variables; the latter two are aspects of the celebrated *measure concentration phenomenon*, the standard form of which is more adapted to almost isometric questions than to isomorphic ones.] For many more examples of similar arguments in other contexts see [5].

In all such cases, a natural question is: *Is it possible to give an explicit example, or a derandomized algorithm?* An explicit example must have an explicit reason, and this should presumably be reflected by the presence of some additional structure and a more natural, "nicer" end-product. Even more importantly, if a question is motivated by applications, it is usually imperative that the solution be explicit, or at least easily verifiable. In this section we will sketch several sample contexts when a derandomized proof would be desirable, and describe a few attempts at derandomization (or partial derandomization) of constructions that were originally obtained using probabilistic methods.

Very often the object one constructs (a convex body, a normed space, or a subspace or a quotient) can be fully described by a matrix, which in a probabilistic construction will be random. [This link is even more explicit when the objective is to find an operator.] Since we aim at producing explicit matrices that behave like random ones, one may say that this section is mostly about *pseudorandom matrices*.

4.1. Kashin decompositions and linear vs. quadratic programming. We begin by recalling the following spectacular result motivated by questions in approximation theory and usually referred to as *Kashin decomposition* (see [45], [96], [102], [84])

Theorem 4.1 (Kashin decomposition). *Given $m = 2n \in 2\mathbb{N}$, there exist two orthogonal m-dimensional subspaces $E_1, E_2 \subset \mathbb{R}^m$ such that*

$$\frac{1}{8}\|x\|_2 \le \frac{1}{\sqrt{m}}\|x\|_1 \le \|x\|_2 \quad \textit{for all } x \in E_i,\ i = 1, 2. \tag{5}$$

In other words, the space ℓ_1^{2n} is an orthogonal (in the ℓ_2^{2n} sense) sum of two *nearly Euclidean* subspaces. The existence of such a decomposition was surprising because, as is easily seen, on the *entire* space $\mathbb{R}^m$, the ratio between the ℓ_1 and ℓ_2 norms varies between 1 and $\sqrt{m}$ (in fact, the Banach–Mazur distance between ℓ_1^m and ℓ_2^m equals $\sqrt{m}$). [See [84], p. 95, for an exposition of the equally striking infinite-dimensional analogue due to Krivine and independently to Kashin.] A slightly different form of the theorem (the original one) asserts the existence of a matrix $V \in O(n)$ such that

$$\max\{\|x\|_1, \|Vx\|_1\} \ge c\sqrt{n}\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n$$

the graphs of V and $-V$ yield then a desired decomposition of $\mathbb{R}^m$ (with $\frac{1}{8}$ in (5) replaced by $\frac{c}{2}$). In both formulations the standard arguments yield that, for large n,

the assertion holds for nearly all decompositions $E_1 \oplus E_2$ or, resp., for nearly all $V \in O(n)$ (with respect to the corresponding Haar measure); see [7], [90], [55] and their references for an in-depth discussion of other random models. However, no explicit families of E_1, E_2 or V with $n \to \infty$ are known. This leads naturally to

Problem 4.2. Given $n \in \mathbb{N}$, exhibit an explicit Kashin decomposition of ℓ_1^{2n}.

A formally easier, and perhaps more to the point, as it corresponds to *constructive Dvoretzky theorem* for ℓ_1^m (cf. Section 3.2), is the question (also stated as Problem 7 in [70]) about exhibiting explicit proportional nearly Euclidean subspaces of ℓ_1^m, i.e., subspaces $E \subset \ell_1^m$ with $k := \dim E \geq cm$ and $d(E, \ell_2^k) \leq C$. The best to date result of this nature is due to Rudin [88] and yields merely $k = O(\sqrt{m})$. The construction in [88] was based on finite fields and difference sets (or the so called *finite geometries*), and the topic directly considered was that of exact Λ_p-sets for even integers $p \geq 4$. This leads to another question: finding explicit exact Λ_p-sets for other values of p; for definitions and probabilistic results see [17], [106], [18].

A very interesting result (whose relevance is not completely clear yet) in the direction of Problem 4.2 was obtained in [15], which – in our language – contains a constructive version of the quotient of a subspace Theorem 3.1 for the simplex. [See also [77], where this and many more related issues are discussed.]

Theorem 4.3. *Given $n \in \mathbb{N}$, there exists a set $S \subset \mathbb{R}^n$ which is an explicit affine image of an explicit section of the $5n$-dimensional simplex and which verifies*

$$B_2^n \subset S \subset C B_2^n.$$

Moreover, C can be replaced by $1+\varepsilon$, for $\varepsilon \in (0, 1)$, if we use a simplex of dimension $\geq C_1 n \log(2/\varepsilon)$.

One thus finds an explicit approximate of the n-dimensional Euclidean ball B_2^n "hidden" in the $5n$-dimensional simplex. The original motivation for Theorem 4.3 was approximating *quadratic programming* problems by *linear programming* problems while increasing the size of a problem only moderately. Here n is the size of the original quadratic problem related to the Euclidean ball, or to an ellipsoid. The dimension of the simplex corresponds to the size of the linear problem (its faces represent constraints), with the increase in size related to the number of auxiliary variables. Representing auxiliary variables in terms of the original variables corresponds to a section of the simplex, and the affine image, or projection, corresponds to verifying whether there is a point with certain coordinates pre-assigned which verifies the constraints. Finally, ε is the precision of the approximation. We emphasize the very weak dependence of the increase in dimension on ε; it is more standard in similar statements in geometric functional analysis to have in place of $\log(2/\varepsilon)$ a factor which is a power of ε (we again refer to the article [48] in this collection for a more detailed discussion of the almost isometric theory, where this particular issue more properly belongs). This is another indication of possible advantages of explicit objects over random ones.

It is not clear whether the approach of [15] can be developed to handle the symmetric case corresponding to a constructive Dvoretzky theorem for ℓ_1^m (closely related to Problem 4.2), or even to a constructive version of Theorem 3.1 for that space. In any case, it seems that the more directly relevant point of view is here the dual form of the Dvoretzky theorem, or of the Kashin decomposition: find an explicit projection, or an affine image, of the m-dimensional cube (the unit ball of ℓ_∞^m) which approximates a Euclidean ball of dimension $\approx cm$.

A vaguely similar topic in that it connects algorithmic issues (approximating, this time in the *isomorphic* sense, problems in combinatorial optimization by their *semi-definite relaxations*) with functional analytic phenomena (*Grothendieck-type inequalities* and the geometry of various high dimensional convex sets) has been studied, among others, in [4], [67], [3], [47]; see the first three of these articles and [77], and their references, for the background. The same circle of ideas, related to inhomogeneity of high dimensional cubes and linked to some of the issues discussed in Section 3, led to the solution – in the negative, for large dimensions – of the following (central case of the) well-known problem of Knaster stated in 1946 in the *New Scottish Book* and published in 1947 in [49]: *Given a continuous function on the sphere in $\mathbb{R}^n$ and a configuration of n points on that sphere, is there a rotation of the configuration on which the function is constant?* See [46] for details and for the background. We refer to [39] for improvements yielding (negative) solutions also for moderate dimensions (at this point the answer is unknown for n between 4 and about 60; the answer is affirmative for $n < 4$) and to [71] for a link between Knaster-like statements and precise versions of Dvoretzky theorem. [Some such statements may still be true, see [46].]

4.2. Decreasing randomness and expander graphs. A significant step in the direction of the problems stated or hinted in the preceding subsection was made in [10]. The approach of that paper uses the paradigm introduced earlier in combinatorics and computer science: if we do not know how to completely dispense with randomness in certain construction, let us at least reduce the number of random bits needed to implement the construction; see [2] for an early article in that direction, usually referred to as the *combinatorial derandomization*. As in [2], pseudorandomness is brought in by *pseudorandom expander graphs* based on Kazhdan property T for groups (see [62], [57]). Rather than flipping the coin $2n^2$ times to obtain a $n \times 2n$ matrix of ± 1's, which represents a linear map from $\mathbb{R}^{2n}$ to $\mathbb{R}^n$ whose kernel is typically an n-dimensional subspace of $\mathbb{R}^{2n}$, one identifies the 2^{2n} possible rows of such matrix (i.e., vectors of length $2n$ with ± 1 coordinates) with vertices of an appropriate explicit expander graph, and then decides which vertices/rows to use by performing a random walk on that graph. Obtaining n rows requires n (or $n-1$) steps, for which we need approximately $n \log_2 d$ random bits (where d is the degree of the graph), to which we need to add $2n$ bits for a random choice of the starting point. Specifics depend on a particular problem considered (the technical details are no longer primarily combinatorial, but field specific), for example to obtain a nearly Euclidean n-dimensional subspace of

ℓ_1^{2n} (or a partially derandomized Dvoretzky theorem for that space), d can be chosen to be polynomial in n and so the cost in random bits is of order $n \log n$.

This is an extremely interesting and promising approach. In addition to Dvoretzky theorem for ℓ_1^n, the authors of [10] partially derandomize, among others, the quotient of a subspace Theorem 3.1. Full derandomization does not seem to be possible there since the initial space X is not concrete; alternatively, the hypothesis would need to include conditions on the presentation of X. This example, while pointing to the "more correct" questions that should be asked in certain contexts, also reveals the limitations of the approach. However, progress beyond those limitations may conceivably be possible if one uses the more sophisticated pseudorandom techniques from, say, [58], [89], the contributions whose full implications have not been completely "digested" yet.

It is interesting to note that, in addition to the *classical* theoretical computer science, the same paradigm (from randomizing to partial derandomizing) and the same underlying techniques have been exploited in the *quantum information theory*, see [36], [6]; this circle of ideas also vaguely relates to random codes from Subsection 4.4 below. Thus one may hope that the interaction of asymptotic geometric analysis and the quantum theory expands beyond the initial encounters such as [12], [100].

4.3. Reducibility of matrices and the property τ. An $n \times n$ matrix M is said to be reducible if, in some orthonormal basis, it can be written as a block matrix

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix},$$

where M_1 and M_2 are square matrices of sizes which are (necessarily) between 1 and $n-1$. This is equivalent to M commuting with a nontrivial orthogonal projection. Based on an analysis of a large class of natural examples it was suggested around 1980 that, as n increases to ∞, the reducible matrices may become more and more dense in the space of all $n \times n$ matrices. A confirmation of this fact from "experimental mathematics" would have had interesting consequences in the theory of quasidiagonal operators, and (likely) some useful implications for numerical linear algebra. However, this hope was soon laid to rest by the following result [37]

Theorem 4.4. *There is a computable constant $c > 0$ such that for every $n \geq 2$ there is an $n \times n$ (real or complex) matrix of norm one which cannot be approximated within c by a reducible matrix.*

The argument given in [37] was non-constructive, the "poorly" reducible matrices being random (albeit of a somewhat special form), and the value of c obtained there was of order 10^{-7}. A construction yielding explicit pseudorandom matrices which are poorly approximable by reducible matrices was given recently in [16] (cf. [99], [109]). The construction in [16] is based on property τ from representation theory.

[A preprint containing a simpler, but weaker, version using Kazhdan's property T was circulated among some specialists in 2002.] The construction depends on noting that a unitary representation $g \to \pi(g)$ on $\mathbb{C}^n$ is irreducible iff the adjoint representation $g \to \mathrm{Ad}_\pi(g)$, defined by $\mathrm{Ad}_\pi(g)(X) := \pi(g)X\pi(g)^*$, does not have non-trivial fixed points when restricted to (the invariant subspace of) trace zero $n \times n$ matrices. On the other hand, the property T (or τ) of a group G says, roughly, that every failure of a unitary representation ρ of G to have non-trivial fixed points can be witnessed in a uniform way on the finite set $S = \{\rho(g_1), \ldots, \rho(g_k)\}$, where $g_1, \ldots, g_k$ are generators of G (independent of ρ). Careful but elementary calculations involving various matrix ideal norms show then that irreducibility of ρ can be likewise (uniformly) witnessed on S, and the argument is concluded, as in [37], by producing an appropriate block matrix some of whose entries are elements of S. The key point in the argument is that k and the estimates quantifying irreducibility and lack of non-trivial fixed points are independent of the dimension of the representation (of course, to begin with, we need to choose a group which – in addition to possessing property τ – has many finite dimensional representations; e.g., $\mathrm{SL}_2(\mathbb{Z})$ fulfills this role well). Again, as a bonus, we get a constant c which is better than that in [37] by several orders of magnitude.

While this argument appears to be tightly connected to the Hilbert space structure and, accordingly, not immediately applicable to our more general setting of normed spaces, it is conceivable that (for example) by considering specific instances of the principle that is behind the construction, and by appealing in a deeper way to their structure, one may obtain pseudorandom matrices that are of relevance to some of the questions suggested elsewhere in this article. [The fact that Hilbert spaces are *the* setting for property T or τ is not disqualifying *per se*; in fact, constructions of, say, random bodies typically appeal to Euclidean structures of the underlying spaces by working, e.g., with Gaussian measures.]

4.4. Random linear codes and other topics. Other situations calling for pseudorandom models that have been mentioned in this article are Gluskin-type random Banach spaces [31], [32], [97], [98], [59], some of which are implicit in Section 3.1, or the spaces exemplifying the saturation phenomenon from Section 3.3. We point out that while the latter spaces depend on the initial, *a priori* arbitrary lower dimensional space W with which we saturate them, the dependence is very canonical. Indeed, what really counts is the arrangement of a finite family of lower dimensional subspaces in the larger space, just as Gluskin-type spaces exploited, in a sense, arrangements of finite sets of points. A construction that comes to mind here is [52], which, in particular, contains a successful derandomization of the example of a space with several extremal parameters, including the so called unconditional basis constant, given previously in [28] as an application of non-constructive Kashin decomposition (Theorem 4.1). However, the approach of [52] is based on spherical codes constructed via finite geometries and so its applicability seems somewhat limited, cf. our discussion of [88] in the paragraph following Theorem 4.1.

Another, quite different question is related to modeling free random variables in

free probability (see [110], [112], [111], [38]) with independent random matrices of increasing size. Due to the apparent central role of the idea of *freeness* in the subject of random matrices it would be very interesting to have also sufficiently canonical pseudorandom models (we note that there exist here constructions based on Clifford matrices [110], [94] which are, however, not fully satisfactory).

We conclude by describing briefly one more development that occurred recently on the border of high-dimensional convexity and computer science and which concerns self-correcting linear codes. The context is roughly as follows. We want to transmit a signal which is a vector $x \in \mathbb{R}^n$. Since some coordinates may get corrupted in the transmission, we introduce some *redundancy* by transmitting instead the vector $y = Ax \in \mathbb{R}^N$, where A is an $N \times n$ matrix independent of x and N is larger, but *not much larger* than n. We then hope that if not too many of the coordinates of y get corrupted in the transmission, then we will be able to recover, in a robust way, the original signal x. In this context, some specific efficient strategies (based on linear programming) for recovery of the original signal along this line were proposed by Donoho and his collaborators (see, e.g., [25] and references in [22]), and existence of very efficient codes, with redundancy close to the theoretical minimum (which depends, of course, on the reliability of the transmission channel) was shown in [23], [87], [22]. However, in a twist which is reminiscent of the classical random *Shannon codes* [93], in the most efficient encoding schemes the matrix A is not explicit! This is not the worst possible scenario since once an appropriate A (of given size) is found in the pre-processing stage, it subsequently can be repeatedly used to encode *all possible* signals $x \in \mathbb{R}^n$. However, in spite of some promising leads, fully satisfactory constructive and algorithmically efficient methods for producing large encoding matrices are still missing here.

We refer the reader to [21] for more background on the topic mentioned above and for other information/communication theory problems that have a similar flavor, and to [66] for a study of linear encoding for random models more general than that of [23], [87], [22] (but still employing the same setup involving vectors from $\mathbb{R}^n$ and $\mathbb{R}^N$).

References

[1] Ageev, S. M., Bogatyĭ, S. A., Repovsh, D., The Banach-Mazur compactum is an Aleksandrov compactification of a Q-manifold. *Mat. Zametki* **76** (1) (2004), 3–10; English transl. *Math. Notes* **76** (1–2) (2004), 3–9.

[2] Ajtai, M., Komlós, J., Szemerédi, E., Deterministic simulation in logspace. In *Proceedings of the 19th Annual ACM Conference on Theory of Computing*, ACM Press, New York 1987, 132–140.

[3] Alon, N., K. Makarychev, K., Makarychev, Y., Naor, A., Quadratic forms on graphs. *Invent. Math.* **163** (3) (2006), 499–522.

[4] Alon, N., Naor, A., Approximating the Cut-Norm via Grothendieck's Inequality. *SIAM J. Computing* **35** (4) (2006), 787–803.

[5] Alon, N., Spencer, J. H., *The probabilistic method. With an appendix on the life and work of Paul Erdös.* 2nd edition, Wiley-Intersci. Ser. Discrete Math. Optim., Wiley-Interscience, New York 2000.

[6] Ambainis, A., Smith, A., Small Pseudo-random Families of Matrices: Derandomizing Approximate Quantum Encryption. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*. Lecture Notes in Comput. Sci. 3122, Springer-Verlag, Berlin 2004, 249–260.

[7] Anderson, G. W., Integral Kašin splittings. *Israel J. Math.* **138** (2003), 139–156.

[8] Artstein, S., Milman, V. D., Szarek, S. J., Duality of Metric Entropy. *Ann. of Math.* (2) **159** (3) (2004), 1313–1328.

[9] Artstein, S., Milman, V. D., Szarek, S. J., Tomczak-Jaegermann, N., On convexified packing and entropy duality. *Geom. Funct. Anal.* **14** (5) (2004), 1134–1141.

[10] Artstein-Avidan, S., Milman, V. D., Logarithmic reduction of the level of randomness in some probabilistic constructions. *J. Funct. Anal.* **235** (1) (2006), 297–329.

[11] Asplund, E., Comparison between plane symmetric convex bodies and parallelograms. *Math. Scand.* **8** (1960), 171–180.

[12] Aubrun, G., Szarek, S. J., Tensor products of convex sets and the volume of separable states on N qudits. *Phys. Rev. A.* **73** (2006), 022109.

[13] Banaszczyk, W., Litvak, A. E., Pajor A., Szarek, S. J., The flatness theorem for non-symmetric convex bodies via the local theory of Banach spaces. *Math. Oper. Res.* **24** (3) (1999), 728–750.

[14] Bárány, I., Random points, convex bodies, lattices. In *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing 2002, 527–535.

[15] Ben-Tal, A., Nemirovski, A., On polyhedral approximations of the second-order cone. *Math. Oper. Res.* **26** (2) (2001), 193–205.

[16] Benveniste, E. J., Szarek, S. J., Property T, property τ, and irreducibility of matrices. In preparation.

[17] Bourgain, J., Bounded orthogonal systems and the $\Lambda(p)$-set problem. *Acta Math.* **162** (1989), 227–245.

[18] Bourgain, J., Λ_p-sets in analysis: results, problems and related aspects. In [41], Vol. 1, 195–232.

[19] Bourgain, J., Pajor, A., Szarek, S. J., Tomczak-Jaegermann, N., On the duality problem for entropy numbers of operators. In *Geometric aspects of functional analysis* (1987–88). Lecture Notes in Math. 1376, Springer-Verlag, Berlin, New York 1989, 50–63.

[20] Bourgain, J., Szarek, S.J., The Banach-Mazur distance to the cube and the Dvoretzky-Rogers factorization. *Israel J. Math.* **62** (2) (1988), 169–180.

[21] Candès, E. J., Compressive sampling. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume III, EMS Publishing House, Zürich 2006, 1433–1452.

[22] Candès, E., Rudelson, M., Vershynin, R., Tao, T., Error correction via Linear Programming, FOCS 2005 (46th Annual Symposium on Foundations of Computer Science), 295–308.

[23] Candès, E. J., Tao., T., Decoding by linear programming. Available on the arXiv preprint server: math.MG/0502327

[24] Diestel, J., Jarchow, H., Pietsch, A., Operator ideals. In [41], Vol. 1, 437–496; Addenda and corrigenda: Vol. 2, 1821.

[25] Donoho, D. L., Huo., X., Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), 2845–2862.

[26] Dudley, R. M., The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1** (1967) 290–330.

[27] Dvoretzky, A., Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces* (Jerusalem, 1960), Jerusalem Academic Press, Jerusalem; Pergamon Press, Oxford, 1961, 123–160.

[28] Figiel, T., Kwapień, S., Pełczyński, A., Sharp estimates for the constants of local unconditional structure of Minkowski spaces. *Bull. Acad. Polon. Sci.* (Sér. Sci. Math. Astronom. Phys.) **25** (12) (1977), 1221–1226.

[29] Giannopoulos, A. A., A note on the Banach-Mazur distance to the cube. In *Geometric aspects of functional analysis* (Israel, 1992–1994), Oper. Theory Adv. Appl. 77, Birkhäuser, Basel 1995, 67–73.

[30] Giannopoulos, A. A., Milman, V. D., Euclidean structure in finite dimensional normed spaces. In [41], Vol. 1, 707–779.

[31] Gluskin, E. D., The diameter of Minkowski compactum roughly equals to n. *Funct. Anal. Appl.* **15** (1981), 57–58.

[32] Gluskin, E. D., Finite-dimensional analogues of spaces without a basis. *Dokl. Akad. Nauk SSSR* **261** (5) (1981), 1046–1050; English transl. *Soviet Math. Dokl.* **24** (3) (1981), 641–644.

[33] Gordon, Y., König, H., Schütt, C., Geometric and probabilistic estimates for entropy and approximation numbers of operators. *J. Approx. Theory* **49** (3) (1987), 219–239.

[34] Gordon, Y., Litvak, A.E., Meyer, M., Pajor, A., John's Decomposition in the General Case and Applications. *J. Differential Geom.* **68** (1) (2004), 99–119.

[35] Grötschel, M., Lovász, L., Schrijver, A., *Geometric algorithms and combinatorial optimization.* Algorithms and Combinatorics 2, Springer-Verlag, Berlin 1993.

[36] Hayden, P., Leung, D., Shor, P. W., Winter, A., Randomizing quantum states: Constructions and applications. *Comm. Math. Phys.* **250** (2) (2004), 371–391.

[37] Herrero, D., Szarek, S. J., How well can an $n \times n$ matrix be approximated by reducible ones? *Duke Math. J.* **53** (1986), 233–248.

[38] Hiai, F., Petz, D., *The semicircle law, free random variables and entropy.* Math. Surveys Monogr. 77, Amer. Math. Soc., Providence, RI, 2000.

[39] Hinrichs, A., Richter, C., The Knaster problem: more counterexamples. *Israel J. Math.* **145** (2005), 311–324.

[40] John, F., Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*. Interscience Publishers, Inc., New York, NY, 1948, 187–204.

[41] *Handbook of the geometry of Banach spaces* (ed. by W. B. Johnson and J. Lindenstrauss). North-Holland, Amsterdam, Vol. 1, 2001 and Vol. 2, 2003.

[42] Johnson, W. B., Lindenstrauss, J., Basic concepts in the geometry of Banach spaces. In [41], Vol. 1, 1–84.

[43] Johnson, W. B., Schechtman, G. Finite dimensional subspaces of L_p. In [41], Vol. 1, 837–870.

[44] Kadets, M. Ĭ., Snobar, M. G., Certain functionals on the Minkowski compactum. *Mat. Zametki* **10** (1971), 453–457; English transl. *Math. Notes* **10** (1971), 694–696.

[45] Kashin, B. S., The widths of certain finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR* (Ser. Mat.) **41** (2) (1977), 334–351, 478; English transl. *Math. USSR-Izv.* **11** (2) (1977), 317–333.

[46] Kashin, B.S., Szarek, S.J., The Knaster problem and the geometry of high-dimensional cubes. *C. R. Math. Acad. Sci. Paris* **336** (11) (2003), 931–936.

[47] Kashin, B.S., Szarek, S.J., On the Gram Matrices of Systems of Uniformly Bounded Functions. *Proc. Steklov Inst. Math.* **243** (2003), 227–233

[48] Klartag, B., Isomorphic and almost-isometric problems in high dimensional convex geometry. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 1547–1562.

[49] Knaster, B., Problem 4. *Colloq. Math.* **30** (1947), 30–31.

[50] Koldobsky, A., König, H., Aspects of the isometric theory of Banach spaces. In [41], Vol. 1, 899–939.

[51] König, H., Milman, V. D., On the covering numbers of convex bodies. In *Geometric aspects of functional analysis* (1985–86), Lecture Notes in Math. 1267, Springer-Verlag, Berlin, New York 1987, 82–95.

[52] König, H., Tomczak-Jaegermann, N., Bounds for projection constants and 1-summing norms. *Trans. Amer. Math. Soc.* **320** (2) (1990), 799–823.

[53] Kuelbs, J., W. V. Li, Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** (1) (1993), 133–157.

[54] Lassak, M., Approximation of convex bodies by centrally symmetric bodies. *Geom. Dedicata* **72** (1) (1998), 63–68.

[55] Litvak, A. E., Pajor, A., Rudelson, M., Tomczak-Jaegermann, N., Vershynin, R., Euclidean embeddings in spaces of finite volume ratio via random matrices. *J. Reine Angew. Math.* **589** (2005), 1–19.

[56] Ledoux, M., Zinn, J., Probabilistic limit theorems in the setting of Banach spaces. In [41], Vol. 2, 1177–1200.

[57] Lubotzky, A. *Discrete groups, expanding graphs and invariant measures. With an appendix by Jonathan D. Rogawski.* Progr. Math. 125, Birkhäuser Verlag, Basel 1994.

[58] Lubotzky, A., Phillips, R., Sarnak, P., Ramanujan graphs. *Combinatorica* **8** (3) (1988), 261–277.

[59] Mankiewicz, P., Tomczak-Jaegermann, N., Quotients of finite-dimensional Banach spaces; random phenomena. In [41], Vol. 2, 1201–1246.

[60] Mankiewicz, P., Szarek, S. J., Random Banach Spaces. The limitations of the method. *Mathematica* **41** (1994), 239–250; Corrigenda: *Mathematica* **42** (1995), 220–221.

[61] Marcus, M. B., Pisier, G., Characterizations of almost surely continuous p-stable random Fourier series and strongly stationary processes. *Acta Math.* **152** (3–4) (1984), 245–301.

[62] Margulis, G. A., Explicit constructions of expanders. *Problemy Peredači Informacii* **9** (1973), 71–80; English transl. *Problems Inform. Transmission* **9** (1973), 325–332.

[63] Maurey, B., Banach spaces with few operators. In [41], Vol. 2, 1247–1297.

[64] Maurey, B., Type, cotype and K-convexity. In [41], Vol. 2, 1299–1332.

[65] Maurey, B., Pisier, G., Séries de variables aléatoires vectorielles indépendantes et propriétés géométriques des espaces de Banach. *Studia Math.* **58** (1) (1976), 45–90.

[66] Mendelson, S., Pajor, A., Tomczak-Jaegermann, N., Reconstruction and subgaussian operators Available on the arXiv preprint server: math.FA/0506239.

[67] Megretski, A., Relaxation of Quadratic Programs in Operator Theory and System Analysis. In *Systems, Approximation, Singular Integral Operators, and Related Topics* (Bordeaux, 2000), Oper. Theory Adv. Appl. 129, Birkhäuser, Basel 2001, 365–392.

[68] Milman, V. D., A new proof of the theorem of A. Dvoretzky on sections of convex bodies. *Funct. Anal. Appl.* **5** (1971), 28–37.

[69] Milman, V. D., Almost Euclidean quotient spaces of subspaces of a finite-dimensional normed space. *Proc. Amer. Math. Soc.* **94** (3) (1985), 445–449.

[70] Milman, V. D., The concentration phenomenon and linear structure of finite-dimensional normed spaces. In *Proceedings of the International Congress of Mathematicians* (Berkeley, Calif., 1986), Vol. 2, Amer. Math. Soc., Providence, R.I., 1987, 961–975.

[71] Milman, V. D., A few observations on the connections between local theory and some other fields. In *Geometric aspects of functional analysis* (1986/87), Lecture Notes in Math. 1317, Springer-Verlag, Berlin 1988, 283–289.

[72] Milman, V. D., Surprising geometric phenomena in high-dimensional convexity theory. In *European Congress of Mathematics* (Budapest, 1996), Vol. II. Progr. Math. 169, Birkhäuser, Basel 1998, 73–91.

[73] Milman, V. D., Randomness and pattern in convex geometric analysis. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 665–677.

[74] Milman, V. D., Pajor, A., Entropy and asymptotic geometry of non-symmetric convex bodies. *Adv. Math.* **152** (2) (2000), 314–335.

[75] Milman, V. D., Schechtman, G. *Asymptotic theory of finite-dimensional normed spaces. With an appendix by M. Gromov.* Lecture Notes in Math. 1200, Springer-Verlag, Berlin 1986.

[76] Milman, V. D., Szarek, S. J., A geometric approach to duality of metric entropy. *C. R. Acad. Sci. Paris Sér. I Math.* **332** (2) (2001), 157–162.

[77] Nemirovski, A., Advances in convex optimization: conic programming. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume I, EMS Publishing House, Zürich 2006/2007.

[78] Pajor, A., Tomczak-Jaegermann, N., Volume ratio and other s-numbers of operators related to local properties of Banach spaces. *J. Func. Anal.* **87** (2) (1989), 273–293.

[79] Pietsch, A., *Theorie der Operatorenideale* (*Zusammenfassung*). Wissenschaftliche Beiträge der Friedrich-Schiller-Universität Jena, Friedrich-Schiller-Universität, Jena, 1972.

[80] Pisier, G., Un théorème sur les opérateurs linéaires entre espaces de Banach qui se factorisent par un espace de Hilbert. *Ann. Sci. Ecole Norm. Sup.* (4) **13** (1) (1980), 23–43.

[81] Pisier, G., Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse Fonctionnelle*, 1980–1981, Exp. No. V, École Polytechnique, Palaiseau 1981, 13 pp.

[82] Pisier, G., Holomorphic semigroups and the geometry of Banach spaces. *Ann. of Math.* (2) **115** (2) (1982), 375–392.

[83] Pisier, G., Finite rank projections on Banach spaces and a conjecture of Grothendieck. In *Proceedings of the International Congress of Mathematicians* (Warszawa, 1983), Vol. 2, PWN, Warsaw 1984, 1027–1039.

[84] Pisier, G., *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Math, 94, Cambridge University Press, Cambridge 1989.

[85] Pisier, G., A new approach to several results of V. Milman. *J. Reine Angew. Math.* **393** (1989), 115–131.

[86] Rudelson, M., Distances between non-symmetric convex bodies and the MM^*-estimate. *Positivity* **4** (2) (2000), 161–178.

[87] Rudelson, M., Vershynin, R., Geometric approach to error correcting codes and reconstruction of signals. *Internat. Math. Res. Notices* **2005** (64) (2005), 4019–4041.

[88] Rudin, W., Trigonometric series with gaps. *J. Math. Mech.* **9** (1960), 203–227.

[89] Sarnak, P., *Some applications of modular forms.* Cambridge Tracts in Math. 99, Cambridge University Press, Cambridge 1990.

[90] Schechtman, G., Special orthogonal splittings of L_1^{2k}. *Israel J. Math.* **139** (2004), 337–347.

[91] Schütt, C., Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory* **40** (2) (1984), 121–128.

[92] Schütt, C., Werner, E., Polytopes with vertices chosen randomly from the boundary of a convex body. In *Geometric aspects of functional analysis*, Lecture Notes in Math. 1807, Springer-Verlag, Berlin 2003, 241–422.

[93] Shannon, C. E., Weaver, W., *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, Ill., 1949.

[94] Shlyakhtenko, D., Limit distributions of matrices with bosonic and fermionic entries. In *Free probability theory* (Waterloo, ON, 1995), Fields Inst. Commun. 12, Amer. Math. Soc., Providence, RI, 1997, 241–252.

[95] Sudakov, V. N., Gaussian random processes, and measures of solid angles in Hilbert space. *Dokl. Akad. Nauk SSSR* **197** (1971), 43–45; English transl. *Soviet Math. Dokl.* **12** (1971), 412–415.

[96] Szarek, S. J., On Kashin's almost Euclidean orthogonal decomposition of ℓ_1^n. *Bull. Acad. Polon. Sci.* (Sér. Sci. Math. Astronom. Phys.) **26** (8) (1978), 691–694.

[97] Szarek, S. J., The finite-dimensional basis problem with an appendix on nets of Grassmann manifolds. *Acta Math.* **151** (3–4) (1983), 153–179.

[98] Szarek, S. J., Spaces with large distance to ℓ_∞^n and random matrices. *Amer. J. Math.* **112** (6) (1990), 899–942.

[99] Szarek, S. J., An exotic quasidiagonal operator. *J. Funct. Anal.* **89** (1990), 274–290.

[100] Szarek, S. J., The volume of separable states is super-doubly-exponentially small in the number of qubits. *Phys. Rev. A* **72** (2005), 032304.

[101] Szarek, S. J., Talagrand, M., An "isomorphic" version of the Sauer-Shelah lemma and the Banach-Mazur distance to the cube. In *Geometric aspects of functional analysis* (1987–88), Lecture Notes in Math. 1376, Springer-Verlag, Berlin 1989, 105–112.

[102] Szarek, S. J., Tomczak-Jaegermann, N., On nearly Euclidean decomposition for some classes of Banach spaces. *Compositio Math.* **40** (3) (1980), 367–385.

[103] Szarek, S. J., Tomczak-Jaegermann, N., Saturating constructions for normed spaces. *Geom. Funct. Anal.* **14** (6) (2004), 1352–1375.

[104] Szarek, S. J., Tomczak-Jaegermann, N., Saturating constructions for normed spaces II. *J. Funct. Anal.* **221** (2) (2005), 407–438.

[105] Szarek, S. J., Tomczak-Jaegermann, N., On the nontrivial projection problem. In preparation.

[106] Talagrand, M., Sections of smooth convex bodies via majorizing measures. *Acta Math.* **175** (1995), 273–300.

[107] Tomczak-Jaegermann, N., Dualité des nombres d'entropie pour des opérateurs à valeurs dans un espace de Hilbert. (French) *C. R. Acad. Sci. Paris Sér. I Math.* **305** (7) (1987), 299–301.

[108] Tomczak-Jaegermann, N., *Banach-Mazur distances and finite-dimensional operator ideals.* Pitman Monogr. Surveys Pure Appl. Math. 38, Longman, Harlow; Wiley, New York 1989.

[109] Voiculescu, D., Property T and approximations of operators. *Bull. London Math. Soc.* **22** (1990), 25–30.

[110] Voiculescu, D., Limit laws for random matrices and free products. *Invent. Math.* **104** (1991), 201–220.

[111] Voiculescu, D., Free probability theory: random matrices and von Neumann algebras. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 227–241.

[112] Voiculescu, D., Dykema, K., Nica, A., *Free random variables. A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups.* CRM Monogr. Ser. 1, Amer. Math. Soc., Providence, RI, 1992.

Case Western Reserve University, Department of Mathematics, Cleveland,
Ohio 44106-7058, U.S.A.
and
Université Pierre et Marie Curie-Paris 6, UMR 7586-Institut de Mathématiques,
Analyse Fonctionnelle, BC 186, 75252 Paris, France
E-mail: szarek@case.edu

Higher index theory of elliptic operators and geometry of groups

Guoliang Yu*

Abstract. The Atiyah–Singer index theorem has been vastly generalized to higher index theory for elliptic operators in the context of noncommutative geometry. Higher index theory has important applications to problems in differential topology and differential geometry such as the Novikov Conjecture on homotopy invariance of higher signatures and the existence problem of Riemannian metrics with positive scalar curvature. In this article, I will give a survey on recent development of higher index theory, its applications, and its fascinating connection to the geometry of groups and metric spaces.

Mathematics Subject Classification (2000). Primary 46L87, 58J20; Secondary 46L80, 20F65.

Keywords. Noncommutative geometry, index theory, elliptic operator, K-theory, operator algebras, metric geometry, geometric group theory.

1. Introduction

An elliptic differential operator D on a compact manifold M is Fredholm in the sense that the kernel and cokernel of D are finite dimensional and the image of D is closed. The Fredholm index of D is defined to be

$$\operatorname{index}(D) = \dim(\ker D) - \dim(\operatorname{coker} D).$$

Roughly speaking, $\operatorname{index}(D)$ measures the size of the solution space of a certain system of differential equations associated to D. The Fredholm index has the following fundamental properties: (1) it is an obstruction to invertibility of D; (2) it is invariant under homotopy equivalence. These are essential properties for the purpose of applications. The celebrated Atiyah–Singer index theorem computes the Fredholm index of elliptic differential operators on compact manifolds and has important applications [4].

Elliptic differential operators on noncompact manifolds are in general not Fredholm in the usual sense but Fredholm in a generalized sense. The generalized Fredholm index for such operators is called the higher index. Higher index theories for elliptic operators in various noncompact situations have been successfully developed

*The author is partially supported by NSF and NSFC.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

by Kasparov [51], [52], Mishchenko–Fomenko [60], Baum–Connes [6], Connes–Skandalis [23], Connes–Moscovici [22], and Roe [67]. In this survey, we will focus on higher index theory in the following two cases: (1) noncompact manifolds with proper and free cocompact actions of discrete groups, (2) noncompact complete Riemannian manifolds.

In the case of a manifold $\tilde{M}$ with a proper and free cocompact action of a discrete group Γ, let M be the compact quotient $\tilde{M}/\Gamma$, let D be an elliptic differential operator on M and let $\tilde{D}$ be the lifting of D to $\tilde{M}$. $\tilde{D}$ is a generalized Fredholm operator and its generalized Fredholm index is an element in the K-theory of the reduced group C^*-algebra $C^*_r(\Gamma)$ [51], [52], [60], [6]. The Fredholm index of D is essentially the 0-dimensional information of the generalized Fredholm index of $\tilde{D}$. For this reason, the generalized Fredholm index of $\tilde{D}$ is called the higher index of $\tilde{D}$. The higher index of $\tilde{D}$, denoted by H-index$(\tilde{D})$, has properties similar to that of the classical Fredholm index: (1) H-index$(\tilde{D})$ is an obstruction to invertibility of $\tilde{D}$; (2) H-index$(\tilde{D})$ is invariant under homotopy equivalence. These properties are crucial for purpose of applications. For example, property (1) implies that if H-index$(\tilde{D})$ is non-zero for the Dirac operator, then the manifold M can not carry a Riemannian metric of positive scalar curvature (as a consequence of the Lichnerowicz formula, positive scalar curvature implies that the Dirac operator $\tilde{D}$ is invertible) [69]. The Baum–Connes Conjecture provides an algorithm to compute the higher index of $\tilde{D}$. The Baum–Connes Conjecture was proved by Higson–Kasparov when Γ has the Haagerup property (e.g. all amenable groups) [43] and by Lafforgue for a large class of groups with Property T [56]. I should also mention important work by Puschnigg [66] and Chabert–Echterhoff–Nest [14]. The general problem of computing the higher index of $\tilde{D}$ is still wide open. However, for the purpose of applications to geometry and topology, it is often enough to determine when the higher index is non-zero. The Strong Novikov Conjecture is an algorithm of determining non-vanishing of the higher index. Currently much more is known about this conjecture than the Baum–Connes Conjecture. In this article, we will focus on the recent development of the Strong Novikov Conjecture.

In the case of a general noncompact complete Riemannian manifold M, Roe has introduced a higher index theory for elliptic differential operators on M [67]. The Coarse Baum–Connes Conjecture is an algorithm to compute the higher index of elliptic differential operators on noncompact complete Riemannian manifolds. This conjecture has been proved for a large class of interesting spaces. In general, there are counter-examples to the Coarse Baum–Connes Conjecture [80], [44]. The Coarse Strong Novikov Conjecture is an algorithm of determining non-vanishing of the higher index. There is an unbounded geometry counter-example to this conjecture [80]. This conjecture is still open for spaces with bounded geometry. In this article, we will also report recent progress on the Coarse Baum–Connes Conjecture and the Coarse Strong Novikov Conjecture.

There is a beautiful link between higher index theory and a certain aspect of metric geometry. At this stage, this part of metric geometry is mostly uncharted territory.

One purpose of this survey is to advertise this aspect of metric geometry.

We remark that all manifolds in this article are smooth.

2. Higher index theory of elliptic operators

In this section we briefly review the higher index theory of elliptic operators. The first part of this section is devoted to higher index theory for noncompact manifolds with a proper and free cocompact action of a discrete group [51], [52], [60], [6]. In the second part of this section, we discuss higher index theory for noncompact complete Riemannian manifolds [67].

2.1. Higher index theory and discrete groups. Let Γ be a discrete group acting properly and freely on a manifold $\tilde{M}$ with compact quotient $M = \tilde{M}/\Gamma$.

We first recall the concept of the reduced group C^*-algebra for any countable discrete group Γ.

Let $l^2(\Gamma)$ be the Hilbert space defined by

$$l^2(\Gamma) = \{\xi : \Gamma \to \mathbb{C} \mid \textstyle\sum_{\gamma\in\Gamma} |\xi(\gamma)|^2 < \infty\}.$$

For each $g \in \Gamma$ we define a unitary operator U_g acting on $l^2(\Gamma)$ by

$$(U_g\xi)(\gamma) = \xi(g^{-1}\gamma)$$

for all $\xi \in l^2(\Gamma)$ and $\gamma \in \Gamma$.

We define the group algebra $\mathbb{C}\Gamma$ by

$$\mathbb{C}\Gamma = \{ \textstyle\sum_{g\in\Gamma} c_g U_g : c_g \in \mathbb{C}\},$$

where $\sum_{g\in\Gamma} c_g U_g$ is a finite sum. Observe that $\mathbb{C}\Gamma$ is an algebra over $\mathbb{C}$.

Definition 2.1. The *reduced* group C^*-algebra $C_r^*(\Gamma)$ is the closure of $\mathbb{C}\Gamma$ under operator norm.

In general, $C_r^*(\Gamma)$ is a highly noncommutative C^*-algebra and is a typical example of a "noncommutative space" in Connes' noncommutative geometry [16].

Let D be an elliptic differential operator on the compact manifold M. Let $\tilde{D}$ be the lifting of D to $\tilde{M}$. Recall that a classical theorem in functional analysis says that an operator is Fredholm if and only if it is invertible modulo K, the algebra of all compact operators. $\tilde{D}$ is in general not Fredholm. However, $\tilde{D}$ is a generalized Fredholm operator in the sense that $\tilde{D}$ is invertible modulo $C_r^*(\Gamma) \otimes K$. By a standard procedure in K-theory, one can define the higher index of $\tilde{D}$, denoted by H-index$(\tilde{D})$, as an element of the K-group $K_0(C_r^*(\Gamma))$. When D is a self-adjoint elliptic differential operator, we can define the higher index of $\tilde{D}$, denoted by H-index$(\tilde{D})$, as an element of the K-group $K_1(C_r^*(\Gamma))$. The higher index of $\tilde{D}$ has the following important

properties: (1) it is an obstruction to invertibility of $\tilde{D}$, (2) it is invariant under homotopy equivalence of $\tilde{D}$.

Let $\mathrm{tr}\colon C_r^*(\Gamma) \to \mathbb{C}$ be the canonical trace defined by

$$\mathrm{tr}(T) = \langle T\delta_e, \delta_e \rangle$$

for every $T \in C_r^*(\Gamma)$, where $\delta_e \in l^2(\Gamma)$ is the Dirac function at the identity element e of the group Γ. This trace induces a homomorphism

$$\mathrm{tr}_*\colon K_0(C_r^*(\Gamma)) \to \mathbb{C}.$$

Atiyah's L^2-index theorem in [3] implies the following identity:

$$\mathrm{tr}_*(H\text{-index}(\tilde{D})) = \mathrm{index}(D).$$

It follows that the Fredholm index of D is the 0-dimensional information of the higher index of $\tilde{D}$.

Let Z be a locally compact space with a proper cocompact Γ-action. The Γ-equivariant K-homology group $K_0^\Gamma(Z)$ is generated by all Γ-invariant "abstract elliptic operators" on Z [51], [52]. Similarly, the Γ-equivariant K-homology group $K_1^\Gamma(Z)$ is generated by all Γ-invariant self-adjoint "abstract elliptic operators" on Z.

Let $\mathcal{E}\Gamma$ be the universal space for proper Γ-actions [7]. The Γ-equivariant K-homology group $K_*^\Gamma(\mathcal{E}\Gamma)$ is defined to be the inductive limit

$$\lim_{Z \subseteq \mathcal{E}\Gamma} K_*^\Gamma(Z),$$

where the inductive limit is taken over all Γ-invariant and Γ-cocompact subsets of $\mathcal{E}\Gamma$.

The Baum–Connes map μ associates each Γ-invariant "abstract elliptic operator" to its higher index:

$$\mu\colon K_*^\Gamma(\mathcal{E}\Gamma) \to K_*(C_r^*(\Gamma)).$$

Conjecture 2.2 (*The Baum–Connes Conjecture* [6], [7]). Let Γ be a countable discrete group. The Baum–Connes map $\mu\colon K_*^\Gamma(\mathcal{E}\Gamma) \to K_*(C_r^*(\Gamma))$ is an isomorphism for Γ.

Let Γ be a discrete group acting properly and freely on a manifold $\tilde{M}$ with compact quotient $M = \tilde{M}/\Gamma$. Let D be an elliptic differential operator on M and $\tilde{D}$ be its lifting to $\tilde{M}$. Denote by $[\tilde{D}]$ the K-homology class of $\tilde{D}$ in $K_*^\Gamma(\tilde{M})$. By the universality of $\mathcal{E}\Gamma$, there exists a Γ-invariant classifying map $f\colon \tilde{M} \to \mathcal{E}\Gamma$.

We have

$$H\text{-index}(\tilde{D}) = \mu(f_*[\tilde{D}]).$$

It follows that the Baum–Connes Conjecture would reduce the computation of the higher index H-index$(\tilde{D})$ to that of $f_*[\tilde{D}]$ in $K_*^\Gamma(\mathcal{E}\Gamma)$, which is in principle computable.

Conjecture 2.3 (*The Strong Novikov Conjecture*). Let Γ be a countable discrete group. The Baum–Connes map $\mu\colon K_*^\Gamma(\mathcal{E}\Gamma) \to K_*(C_r^*(\Gamma))$ is injective for Γ.

The Strong Novikov Conjecture would reduce the non-vanishing problem of the higher index H-index$(\tilde{D})$ to that of $f_*[\tilde{D}]$ in $K_*^\Gamma(\mathcal{E}\Gamma)$, which is in principle decidable.

If M is an aspherical compact manifold (i.e. its universal cover is contractible), the Strong Novikov Conjecture implies the Gromov–Lawson Conjecture which claims that an aspherical compact manifold can not carry a Riemannian metric with positive scalar curvature [67]. Let Γ be the fundamental group of M and $\tilde{M}$ be the universal cover of M. In this case f_* is an isomorphism and the Strong Novikov Conjecture implies that the higher index of the Dirac operator is non-zero [67]. However, by the Lichnerowicz formula, positive scalar curvature would imply invertibility of the Dirac operator. In general, the Strong Novikov Conjecture implies a stable version of the Gromov–Lawson–Rosenberg Conjecture [73]. This stable version of the Gromov–Lawson–Rosenberg Conjecture provides a complete answer to the question when a compact manifold can stably carry a Riemannian metric with positive scalar curvature.

Another very important corollary of the Strong Novikov Conjecture is the Novikov Conjecture on homotopy invariance of higher signatures. The Novikov Conjecture is a central problem in the classification of higher dimensional compact manifolds (see [32] for a historic account of the Novikov Conjecture). In the case of aspherical compact manifolds, the Novikov Conjecture states that rational Pontryagin classes are homotopy invariants. In this case the Novikov Conjecture can be considered as an infinitesimal version of the Borel Conjecture, which claims that any aspherical compact manifold M is rigid in the sense that if another compact manifold M' is homotopy equivalent to M, then M' is homeomorphic to M. The Strong Novikov Conjecture implies an integral version of the Novikov Conjecture in L-theory after inverting 2 [70]. This integral version of the Novikov Conjecture implies the stable Borel Conjecture which states that if a compact manifold M is aspherical and another compact manifold M' is homotopy equivalent to M, then $M' \times \mathbb{R}^n$ is homeomorphic to $M \times \mathbb{R}^n$ for some large n [31].

2.2. Higher index theory for noncompact complete Riemannian manifolds. We shall briefly review Roe's higher index theory for noncompact complete Riemannian manifolds [67].

We first recall the concept of Roe algebra. Let Γ be a locally finite discrete metric space (recall that a discrete metric space is said to be locally finite if every ball has finitely many elements). Let H be a separable and infinite dimensional Hilbert space. We decompose

$$l^2(\Gamma) \otimes H = \oplus_{\gamma \in \Gamma}(\delta_\gamma \otimes H),$$

where $\delta_\gamma \in l^2(\Gamma)$ is the Dirac function at γ. For each bounded linear operator $T : l^2(\Gamma) \otimes H \to l^2(\Gamma) \otimes H$, we have a corresponding decomposition:

$$T = (T_{x,y})_{x,y \in \Gamma},$$

where $T_{x,y}$ is a bounded linear operator from $\delta_y \otimes H$ to $\delta_x \otimes H$.

Definition 2.4 (Roe [67]). Let Γ be a locally finite discrete metric space.

(1) A bounded linear operator $T: l^2(\Gamma) \otimes H \to l^2(\Gamma) \otimes H$ is said to be *locally compact* if $T_{x,y}$ is compact for all $x, y \in \Gamma$;

(2) A bounded linear operator $T: l^2(\Gamma) \otimes H \to l^2(\Gamma) \otimes H$ is said to have *finite propagation* if there exists $R \geq 0$ such that $T_{x,y} = 0$ for all $x, y \in \Gamma$ satisfying $d(x, y) > R$.

(3) The *Roe algebra* $C^*(\Gamma)$ is defined to be the operator norm closure of all locally compact operators acting on $l^2(\Gamma) \otimes H$ with finite propagation.

An important feature of the Roe algebra is that, up to $*$-isomorphism, it depends only on the quasi-isometry type (or more generally the coarse type) of the locally finite discrete metric space.

If Z is a locally compact metric space, we choose a net Γ in Z and define the Roe algebra $C^*(Z)$ by $C^*(\Gamma)$. Recall that a locally finite discrete subspace $\Gamma \subseteq Z$ is said to be a net if there exists $c \geq 0$ satisfying $d(z, \Gamma) \leq c$ for every $z \in Z$. We observe that the definition of $C^*(Z)$ is independent of the choice of Γ up to $*$-isomorphism.

If M is a noncompact complete Riemannian manifold and D is a geometric elliptic operator on M, then D is a generalized Fredholm operator in the sense that it is invertible modulo $C^*(M)$. One can define a higher index, denoted by H-index(D), as an element of the K-group $K_0(C^*(M))$ [67]. Similarly when D is a self-adjoint geometric elliptic operator on M, one can define a higher index, denoted by H-index(D), as an element of the K-group $K_1(C^*(M))$. This higher index is an obstruction to invertibility and is invariant under homotopy equivalence.

Definition 2.5. Let Γ be a locally finite metric space. For each $d \geq 0$, the *Rips complex* $P_d(\Gamma)$ is the simplicial polyhedron where the set of all vertices is Γ, and a finite subset $\{\gamma_0, \ldots, \gamma_n\} \subseteq \Gamma$ spans a simplex iff $d(\gamma_i, \gamma_j) \leq d$ for all $0 \leq i, j \leq n$.

Let Z be a locally compact space. Recall that the K-homology group $K_0(Z)$ is generated by all "abstract elliptic operators" on Z [51], [52]. Similarly, the K-homology group $K_1(Z)$ is generated by all self-adjoint "abstract elliptic operators" on Z.

Let Γ be a locally finite discrete metric space. The coarse Baum–Connes map μ associates every "abstract elliptic operator" to its higher index:

$$\mu: \lim_{d \to \infty} K_*(P_d(\Gamma)) \to K_*(C^*(\Gamma)).$$

Conjecture 2.6 (*The Coarse Baum–Connes Conjecture* [45], [78]). If Γ is a locally finite discrete metric space, then the coarse Baum–Connes map μ is an isomorphism.

If Γ is a countable discrete group with a length metric and its classifying space $B\Gamma$ has the homotopy type of a finite CW-complex, then the descent principle says that the Coarse Baum–Connes Conjecture for Γ as a metric space implies the Strong Novikov Conjecture for Γ as a group [68].

Recall that a function $l: \Gamma \to [0, \infty)$ is said to be a length function if

(1) $l(g) = 0$ for some $g \in \Gamma$ if and only if g is the identity element;

(2) $l(g) = l(g^{-1})$ for all $g \in \Gamma$ and $l(g_1 g_2) \leq l(g_1) + l(g_2)$ for all g_1 and g_2 in Γ;

(3) l is proper in the sense that $l^{-1}(K)$ is a finite subset of Γ for every compact subset K of $[0, \infty)$.

We remark that such a length function always exists for any countable discrete group. If Γ is a finitely generated group, we can construct a word length for any finite generating set. For any length function l on a countable group Γ, we can associate a length metric d on Γ by: $d(g_1, g_2) = l(g_1^{-1} g_2)$ for all g_1 and g_2 in Γ. We remark that a countable discrete group with a length metric has bounded geometry (recall that a discrete metric space Γ is said to have bounded geometry if for all $r > 0$ there exists $N(r) > 0$ such that the number of elements in every ball with radius r is at most $N(r)$).

A counterexample to the injectivity of the coarse Baum–Connes map is found in [81]. However, this example does not have bounded geometry. A bounded geometry counterexample to surjectivity is found in [44].

The following conjecture is still wide open.

Conjecture 2.7 (*The Coarse Strong Novikov Conjecture*). If Γ is a discrete metric space with bounded geometry, then the coarse Baum–Connes map μ is injective.

If M is a noncompact complete Riemannian manifold with bounded geometry (i.e. M has bounded curvature and positive injectivity radius), then there exists a net Γ in M such that Γ has bounded geometry. Let $\{U_\gamma\}_{\gamma \in \Gamma}$ be an open cover of M such that $\gamma \in U_\gamma$ for each $\gamma \in \Gamma$ and $\{\text{diameter}(U_\gamma)\}_{\gamma \in \Gamma}$ is uniformly bounded. Let $\{\phi_\gamma\}_{\gamma \in \Gamma}$ be a partition of unity subordinate to $\{U_\gamma\}_{\gamma \in \Gamma}$. We define a continuous map $\phi \colon M \to P_d(\Gamma)$ for some large d by

$$\phi(x) = \sum_{\gamma \in \Gamma} \phi_\gamma(x)\gamma$$

for all $x \in M$.

Let D be a geometric elliptic operator on M and $[D]$ be its K-homology class in $K_*(M)$.

We have

$$H\text{-index}(D) = \mu(\phi_*[D]).$$

It follows that the Coarse Strong Novikov Conjecture would reduce the non-vanishing problem of H-index(D) to that of $\phi_*([D])$ in $\lim_{d \to \infty} K_*(P_d(\Gamma))$, which is in principle decidable.

Recall that a Riemannian manifold M is said to be uniformly contractible if for every $r > 0$, there exists $R \geq r$ such that for each $x \in M$, the ball with radius r and center x can be contracted to a point within the ball with radius R and center x. If M is a uniformly contractible Riemannian manifold with bounded geometry, then the Coarse

Strong Novikov Conjecture implies a conjecture of Gromov which states that M can not have uniform positive scalar curvature [36]. In this case ϕ_* is an isomorphism and therefore the Coarse Strong Novikov Conjecture implies that the higher index of the Dirac operator is non-zero. However, by the Lichnerowicz formula, uniform positive scalar curvature would imply invertibility of the Dirac operator.

3. Geometry of groups and metric spaces

In this section we briefly discuss several useful concepts from metric geometry. These concepts play important roles in higher index theory.

The following concept plays an important role in the study of Novikov type conjectures.

Definition 3.1 (Gromov [35]). Let Γ be a metric space; let X be a Banach space. A map $f\colon \Gamma \to X$ is said to be a (*coarse*) *uniform embedding* if there exist non-decreasing functions ρ_1 and ρ_2 from $\mathbb{R}_+ = [0, \infty)$ to $\mathbb{R}$ such that

(1) $\rho_1(d(x, y)) \leq \|f(x) - f(y)\| \leq \rho_2(d(x, y))$ for all $x, y \in \Gamma$;

(2) $\lim_{r\to+\infty} \rho_i(r) = +\infty$ for $i = 1, 2$.

Intuitively (coarse) uniform embeddability of a metric space Γ into a Banach space X means that we can draw a "nice" picture of Γ in X which reflects the large scale geometry of Γ. If Γ is a countable discrete group with a length metric, the concept of (coarse) uniform embeddability of Γ into a Banach space X does not depend on the choice of the length metric.

The following construction implies that every discrete metric space admits a (coarse) uniform embedding into some Banach space, and we need to consider reasonably nice Banach spaces to do anything interesting.

Let Γ be a discrete metric space, let $X = l^\infty(\Gamma)$. Fix $x_0 \in \Gamma$. We define an isometric embedding of Γ into X by

$$(f(x))(\gamma) = d(\gamma, x) - d(\gamma, x_0)$$

for every $x, \gamma \in \Gamma$.

The Hilbert space case is particularly important. Examples of groups which admit (coarse) uniform embedding into Hilbert space include

(1) amenable groups (Bekka–Cherix–Valette [8]);

(2) groups with finite asymptotic dimension (e.g. a certain mapping class groups [9]), more generally groups with polynomial asymptotic dimension growth [26];

(3) hyperbolic groups [71], more generally a class of relatively hyperbolic groups [62], [65], [25];

(4) countable subgroups of any almost connected Lie groups (Guentner–Higson–Weinberger [39]);

(5) Coxeter groups [29], more generally certain diagram groups (including R. Thompson's group F) [30], [12], [1];

(6) semi-direct products of groups of the above types.

In [34], Gong–Yu proved that a noncompact Riemannian manifold with subexponential volume growth admits a (coarse) uniform embedding into Hilbert space.

The following concept provides the most effective method to construct (coarse) uniform embedding into Hilbert space.

Definition 3.2 ([81]). A locally finite discrete metric space Γ is said to have *Property A* if for any $r > 0$, $\varepsilon > 0$, there exist a family of finite subsets $\{A_\gamma\}_{\gamma\in\Gamma}$ of $\Gamma \times \mathbb{N}$ ($\mathbb{N}$ is the set of all natural numbers) such that

(1) $(\gamma, 1) \in A_\gamma$ for all $\gamma \in \Gamma$;

(2) if γ and γ' are two points in Γ satisfying $d(\gamma, \gamma') \leq r$, then

$$\frac{\#(A_\gamma - A_{\gamma'}) + \#(A_{\gamma'} - A_\gamma)}{\#(A_\gamma \cap A_{\gamma'})} < \varepsilon,$$

where $\#S$ is the number of elements in S for any finite set S;

(3) there exists $R > 0$ such that if $(x, m) \in A_\gamma$, $(y, n) \in A_\gamma$ for some $\gamma \in \Gamma$, then $d(x, y) \leq R$.

The set of natural numbers $\mathbb{N}$ in the above definition allows one to count points in Γ with multiplicity. We remark that the concept of Property A is invariant under quasi-isometry (more generally it is invariant under coarse equivalence). If Z is a locally compact metric space, we say that Z has Property A if a net Γ of Z has Property A.

Proposition 3.3 (Yu [81]). *If a locally compact metric space Z has Property A, then Z admits a (coarse) uniform embedding into Hilbert space.*

In the case of a countable discrete group Γ with a length metric, Higson–Roe proved that Γ has Property A if and only if Γ acts amenably on some compact space [46]. Guentner–Kaminker proved that a certain exactness of the reduced group C^*-algebra $C^*_r(\Gamma)$ implies that Γ admits a (coarse) uniform embedding into Hilbert space [40]. Ozawa further proved that exactness of the reduced group C^*-algebra $C^*_r(\Gamma)$ is equivalent to Property A of Γ [63]. In [15], Chen–Wang established a very nice connection between Property A and the ideal structure of the Roe algebra.

Many of the known examples of locally finite discrete metric spaces (coarsely) uniformly embeddable into Hilbert space have Property A. In a recent paper [61], Nowak constructed the first example of a locally finite discrete metric spaces which does not have Property A, but admits a (coarse) uniform embedding into Hilbert space.

Based on ideas of Enflo, Dranishnikov–Gong–Lafforgue–Yu found the first example of a locally finite discrete metric space which does not admit a (coarse) uniform

embedding into Hilbert space [28]. But this example does not have bounded geometry. In [37], Gromov used expanding graphs to construct examples of discrete metric spaces with bounded geometry and also finitely generated groups which do not admit a (coarse) uniform embedding into Hilbert space.

For purpose of the Novikov Conjecture, we need to introduce the following convexity conditions for Banach spaces.

Definition 3.4. Let X be a Banach space.

(1) X is said to be *strictly convex* if

$$\left\| \frac{x+y}{2} \right\| < 1$$

whenever $x,\ y \in S(X)$ and $x \neq y$, where $S(X) = \{x \in X,\ \|x\| = 1\}$;

(2) X is called *uniformly convex* if for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $x,\ y \in S(X)$ and $\|x - y\| \geq \varepsilon$, then

$$\left\| \frac{x+y}{2} \right\| < 1 - \delta.$$

Examples of uniformly convex Banach spaces include l^p, L^p, and C_p for all $1 < p < \infty$, where C_p is the Banach space of all Schatten-p class operators on a Hilbert space, i.e. $C_p = \{T : H \to H \mid \mathrm{tr}(T^*T)^{\frac{p}{2}} < \infty\}$ (H is a Hilbert space).

N. Brown and E. Guentner proved that every countable discrete metric space admits a (coarse) uniform embedding into a strictly convex and reflexive Banach space [11]. W. B. Johnson and N. L. Randrianarivony showed that l^p $(p > 2)$ does not admit a (coarse) uniform embedding into a Hilbert space [50]. More recently, M. Mendel and A. Naor proved that l^p does not admit a (coarse) uniform embedding into l^q if $p > q \geq 2$ [57].

Problem 3.5. Let $p > q \geq 2$. Construct a discrete metric space with bounded geometry which admits a (coarse) uniform embedding into l^p but not into l^q;

Problem 3.6. Let $p > q \geq 2$. Construct a countable discrete group which admits a (coarse) uniform embedding into l^p but not into l^q.

Such a group would be fascinating in the world of group theory since it is neither among Gromov's random groups [37] nor among the familiar class of groups constructed algebraically using amenable groups, hyperbolic groups and linear groups.

Question 3.7. Does every discrete metric space with bounded geometry admit (coarse) uniform embedding into a uniformly convex Banach space?

A negative answer to this question might reveal a geometric phenomenon leading to counter-examples of the Novikov Conjecture. Of course, a positive answer to the

question would imply the Novikov Conjecture for every countable discrete group [55]. We should mention that Ozawa proved that if X is a uniformly convex Banach space with an unconditional basis, then a sequence of expanding graphs does not admit a (coarse) uniform embedding into X [64].

The following concept plays a very useful role in my work with Kasparov on the Baum–Connes Conjecture (in preparation).

Definition 3.8. Let X be a Banach space and Γ be a countable discrete group. An affine and isometric action α of Γ on X is said to be *proper* if

$$\lim_{l(g)\to\infty} \|\alpha(g)\xi\| = \infty$$

for every $\xi \in X$, where l is a length metric on Γ.

Roughly speaking, if a group Γ admits a proper and affine isometric action on X, it means that Γ can be effectively realized as symmetries of the Banach space X. If α is a proper affine isometric action of Γ on a Banach space X, then the map $\gamma \to \alpha(\gamma)0$ is a (coarse) uniform embedding of Γ into X, where 0 is the zero vector in X.

We remark that the concept of proper affine isometric action does not depend on the choice of the length function. If Γ admits a proper isometric affine action on Hilbert space, then Γ is said to satisfy the Haagerup property [41] or to be a-T-menable [35].

The following example shows that every countable discrete group admits a proper affine isometric action on some Banach space.

Let $X = l^\infty(\Gamma)$. Let π be the regular action of Γ on X defined by

$$((\pi(g))\xi)(\gamma) = \xi(g^{-1}\gamma)$$

for every $\xi \in X$, all g and γ in Γ. We define an affine and isometric action α on X by:

$$\alpha(g)\xi = \pi(g)\xi + \pi(g)l - l$$

for every $\xi \in X$ and $g \in \Gamma$, where l is a length function on Γ. It is not difficult to verify that α is a proper action.

It is well known that an infinite Property T group does not admit a proper affine isometric action on Hilbert space. A theorem of Zuk says that most (infinite) hyperbolic groups have Property T and therefore do not admit a proper affine isometric action on Hilbert space [83]. The following result shows the usefulness of considering more general uniformly convex Banach spaces.

Proposition 3.9 ([82]). *If Γ is a hyperbolic group, then there exists $2 \le p < \infty$ such that Γ admits a proper affine isometric action on an l^p-space.*

A very recent result of Bader–Furman–Gelander–Monod says that $\mathrm{SL}(n, \mathbb{Z})$ does not admit a proper affine isometric action on l^p-spaces if $n \ge 3$ [5]. A natural question is the following.

Question 3.10. Does $\mathrm{SL}(n, \mathbb{Z})$ admit a proper affine isometric action on some uniformly convex Banach space for $n \geq 3$?

A positive answer to this question would have interesting applications to the Baum–Connes Conjecture.

4. Main results

In this section, we briefly discuss several recent results in higher index theory and their applications.

Theorem 4.1 ([81]). *Let Γ be a discrete metric space with bounded geometry. If Γ admits a (coarse) uniform embedding into Hilbert space, then the Coarse Baum–Connes Conjecture is true for Γ.*

The special case of finite asymptotic dimension was first proved in [80] by a controlled operator K-theory method.

Corollary 4.2. *Let M be a uniformly contractible Riemannian manifold with bounded geometry. If M admits a (coarse) uniform embedding into Hilbert space, then M can not have uniform positive scalar curvature.*

Theorem 4.3 ([81], [74]). *Let Γ be a countable discrete group. If Γ admits a (coarse) uniform embedding into Hilbert space, then the Strong Novikov Conjecture holds for Γ.*

By the descent principle, Theorem 4.3 follows from Theorem 4.1 under an additional finiteness assumption on the homotopy type of the classifying space $B\Gamma$ [81]. This finiteness assumption was removed in [74] with the help of a technique introduced by Higson [42] and a theorem of Tu [75].

Corollary 4.4. *The Novikov Higher Signature Conjecture is true if the fundamental group admits a (coarse) uniform embedding into Hilbert space.*

By discussions in Section 2, the Novikov Conjecture is essentially a topological "recognition" problem for compact manifolds. Roughly speaking, Corollary 4.4 says if we can draw a "nice" picture of the fundamental group of a compact manifold in Hilbert space, then we can "recognize" the manifold topologically.

The following result is a generalization of the injectivity part of Theorem 4.1.

Theorem 4.5 ([54]). *Let Γ be a discrete metric space with bounded geometry. If Γ admits a (coarse) uniform embedding into a uniformly convex Banach space, then the Coarse Strong Novikov Conjecture is true for Γ.*

Corollary 4.6. *If a uniformly contractible Riemannian manifold M with bounded geometry admits a (coarse) uniform embedding into a uniformly convex Banach space, then M can not have uniform positive scalar curvature.*

The following result is a generalization of Theorem 4.3.

Theorem 4.7 ([55]). *Let Γ be a countable discrete group. If Γ admits a (coarse) uniform embedding into a uniformly convex Banach space, then the Strong Novikov Conjecture is true for Γ.*

A key ingredient in the proofs of Theorems 4.5 and 4.7 is the construction of an infinite dimensional Bott bundle over the Banach space. Uniform convexity is used to show that this Bott bundle is almost flat in a certain Banach sense.

Corollary 4.8. *The Novikov Higher Signature Conjecture is true if the fundamental group admits a (coarse) uniform embedding into a uniformly convex Banach space.*

We end this survey with the following folklore conjecture.

Conjecture 4.9. Let M be a compact manifold and $\mathrm{Diff}(M)$ be the group of all diffeomorphisms of M. If Γ is a countable subgroup of $\mathrm{Diff}(M)$, then Γ admits a uniform embedding into C_p for some $1 < p < \infty$, where p depends on the dimension of M and C_p is the Banach space of all Schatten-p class operators on Hilbert space.

It is an open question whether every countable subgroup of $\mathrm{Diff}(M)$ for a compact manifold M admits a proper affine isometric action on C_p for some $1 < p < \infty$. This question and Conjecture 4.9 are open even for the case of the circle. If Conjecture 4.9 is true, then Theorem 4.7 implies the Strong Novikov Conjecture for every countable subgroup of $\mathrm{Diff}(M)$ for a compact manifold M.

Note added in proof. Y. Kida has proved that a mapping class group admits a (coarse) uniform embedding into Hilbert space in his recent preprint "The mapping class group from the viewpoint of measure equivalence theory". U. Hamenstädt has independently obtained the same result in her recent preprint "Geometry of the mapping class groups I: Boundary amenability".

References

[1] Arzhantseva, G., Guba, V., and Sapir, M., Metrics on diagram groups and uniform embeddings in a Hilbert space. Preprint, 2005.

[2] Atiyah, M. F., Global theory of elliptic operators. In *Proc. Internat. Conf. on Functional Analysis and Related Topics* (Tokyo, 1969), University of Tokyo Press, Tokyo 1970, 21–30.

[3] Atiyah, M. F., Elliptic operators, discrete groups and von Neumann algebras. *Colloque "Analyse et Topologie" en l'Honneur de Henri Cartan* (Orsay, 1974), *Asterisque* **32-33**, Soc. Math. France, Paris 1976, 43–72.

[4] Atiyah, M. F., and Singer, I. M., The index of elliptic operators. I. III. *Ann. of Math.* (2) **87** (1968), 484–530; 546–604.

[5] Bader, U., Furman, A., Gelander, T., and Monod, N., Property (T) and rigidity for actions on Banach spaces. Preprint, 2005.

[6] Baum, P., and Connes, A., K-theory for discrete groups. In *Operator Algebras and Applications,* (ed. b D. Evans and M. Takesaki, editors), Cambridge University Press, Cambridge 1989, 1–20.

[7] Baum, P., Connes, A., and Higson, N., Classifying space for proper actions and K-theory of group C^*-algebras. In *C^*-algebras: 1943–1993* (San Antonio, TX, 1993), Contemp. Math. 167, Amer. Math. Soc., Providence, RI, 1994, 240–291.

[8] Bekka, M. E. B., Cherix, P.-A., and Valette, A., Proper affine isometric actions of amenable groups. In *Novikov conjectures, index theorems and rigidity* (Oberwolfach, 1993), Vol. 2, London Math. Soc. Lecture Note Ser. 227, Cambridge University Press, Cambridge 1995, 1–4.

[9] Bell, G., and Fujiwara, K., The asymptotic dimension of a curve graph is finite. Preprint, 2005.

[10] Block, J., and Weinberger, S., Aperiodic tilings, positive scalar curvature and amenability of spaces. *J. Amer. Math. Soc.* **5** (4) (1992), 907–918.

[11] Brown, N., and Guentner, E., Uniform embedding of bounded geometry spaces into reflexive Banach spaces, *Proc. Amer. Math. Soc.* **133** (7) (2005), 2045–2050 (electronic).

[12] Campbell, S., and Niblo, G. A., Hilbert space compression and exactness of discrete groups. *J. Funct. Anal.* **222** (2) (2005), 292–305.

[13] Carlsson, G., and Pedersen, E. K., Controlled algebra and the Novikov conjectures for K- and L-theory. *Topology* **34** (3) (1995), 731–758.

[14] Chabert, J., Echterhoff, S., and Nest, R., The Connes-Kasparov conjecture for almost connected groups and for linear p-adic groups. *Inst. Hautes Études Sci. Publ. Math.* **97** (2003), 239–278.

[15] Chen, X., and Wang, Q., Ideal structure of uniform Roe algebras of coarse spaces. *J. Funct. Anal.* **216** (1) (2004), 191–211.

[16] Connes, A., *Noncommutative Geometry.* Academic Press, San Diego, CA, 1994.

[17] Connes, A., A survey of foliations and operator algebras. In *Operator algebras and applications* (Kingston, Ont., 1980), Part I, Proc. Sympos. Pure Math. 38, Amer. Math. Soc., Providence, R.I., 1982, 521–628.

[18] Connes, A., Noncommutative differential geometry. *Inst. Hautes Études Sci. Publ. Math.* **62** (1985), 257–360.

[19] Connes, A., Cyclic cohomology and transverse fundamental class of a foliation. In *Geometric Methods in Operator Algebras,* (ed. by H. Araki and E. G. Effros), Pitman Res. Notes Math. 123, Longman Sci. Tech., Harlow 1986, 52–144.

[20] Connes, A., Gromov, M., and Moscovici, H., Conjecture de Novikov et fibrés presque plats. *C. R. Acad. Sci. Paris Ser. I Math.* **310** (5) (1990), 273–277.

[21] Connes, A., Gromov, M., and Moscovici, H., Group cohomology with Lipschitz control and higher signatures. *Geom. Funct. Anal.* **3** (1993), 1–78.

[22] Connes, A., and Moscovici, H., Cyclic cohomology, the Novikov conjecture and hyperbolic groups. *Topology* **29** (1990), 345–388.

[23] Connes, A., and Skandalis, G., The longitudinal index theorem for foliations. *Publ. Res. Inst. Math. Sci.* **20** (6) (1984), 1139–1183.

[24] Cuntz, J., Noncommutative simplicial complexes and the Baum-Connes conjecture. *Geom. Funct. Anal.* **12** (2) (2002), 307–329.

[25] Dadarlat, M., and Guentner, E., Uniform embeddibilty of relatively hyperbolic groups. Preprint, 2005.

[26] Dranishnikov, A. N., Groups with a polynomial dimension growth. Preprint, 2004.

[27] Dranishnikov, A. N., Ferry, S. C., and Weinberger, S., Large Riemannian manifolds which are flexible. *Ann. of Math.* (2) **157** (3) (2003), 919–938.

[28] Dranishnikov, A. N., Gong, G., Lafforgue, V., and Yu, G., Uniform embeddings into Hilbert space and a question of Gromov. *Canad. Math. Bull.* **45** (1) (2002), 60–70.

[29] Dranishnikov, A., and Januszkiewicz, T., Every Coxeter group acts amenably on a compact space. In *Proceedings of the 1999 Topology and Dynamics Conference* (Salt Lake City, UT), Topology Proc. **24** (1999), 135–141.

[30] Farley, D., Finiteness and CAT(0) properties of diagram groups. *Topology* **42** (5) (2003), 1065–1082.

[31] Farrell, F. T., and Hsiang, W. C., On Novikov's conjecture for nonpositively curved manifolds. I. *Ann. of Math.* (2) **113** (1) (1981), 199–209.

[32] Ferry, S. C., Ranicki, A., and Rosenberg, J., A history and survey of the Novikov conjecture. In *Novikov conjectures, index theorems and rigidity,* (Oberwolfach, 1993), Vol. 1, London Math. Soc. Lecture Note Ser. 226, Cambridge University Press, Cambridge 1995, 7–66.

[33] Ferry, S. C., and Weinberger, S., A coarse approach to the Novikov conjecture. In *Novikov conjectures, index theorems and rigidity,* (Oberwolfach, 1993), Vol. 1, London Math. Soc. Lecture Note Ser. 226, Cambridge University Press, Cambridge 1995, 147–163.

[34] Gong, G., and Yu, G., Volume growth and positive scalar curvature. *Geom. Funct. Anal.* **10** (4) (2000), 821–828.

[35] Gromov, M., Asymptotic invariants for infinite groups. In *Geometric Group Theory* (ed. by G. A. Niblo and M. A. Roller), Vol. 2, Cambridge University Press, Cambridge 1993, 1–295.

[36] Gromov, M., Positive curvature, macroscopic dimension, spectral gaps and higher signatures. In *Functional Analysis on the eve of the 21st century,* Vol. 2, Progr. Math. 132, Birkhäuser Boston, Boston, MA, 1996, 1–213.

[37] Gromov, M., Spaces and questions. In *GAFA 2000* (Tel Aviv, 1999), *Geom. Funct. Anal.* 2000, Special Volume, Part I, 118–161.

[38] Gromov, M., and Lawson, B., Positive scalar curvature and the Dirac operator on complete Riemannian manifolds. *Inst. Hautes Études Sci. Publ. Math.* **58** (1983), 83–196.

[39] Guentner, E., Higson, N., and Weinberger, S., The Novikov conjecture for linear groups. *Inst. Hautes Études Sci. Publ. Math.*, to appear.

[40] Guentner, E., and Kaminker, J., Exactness and the Novikov conjecture. *Topology* **41** (2) (2002), 411–418.

[41] Haagerup, U., An example of a nonnuclear C^*-algebra, which has the metric approximation property. *Invent. Math.* **50** (3) (1978/79), 279–293.

[42] Higson, N., Bivariant K-theory and the Novikov conjecture. *Geom. Funct. Anal.* **10** (3) (2000), 563–581.

[43] Higson, N., and Kasparov, G. G., E-theory and KK-theory for groups which act properly and isometrically on Hilbert space. *Invent. Math.* **144** (1) (2001), 23–74.

[44] Higson, N., Lafforgue, V., and Skandalis, G., Counterexamples to the Baum-Connes conjecture. *Geom. Funct. Anal.* **12** (2) (2002), 330–354.

[45] Higson, N., and Roe, J., On the coarse Baum–Connes conjecture. In *Novikov Conjectures, Index Theorems and Rigidity* (ed. by S. Ferry, A. Ranicki and J. Rosenberg), Vol. 2, Cambridge University Press, Cambridge 1995, 227–254.

[46] Higson, N., and Roe, J., Amenable group actions and the Novikov conjecture. *J. Reine Angew. Math.* **519** (2000), 143–153.

[47] Higson, N., Roe, J., and Yu, G., A coarse Mayer–Vietoris principle, *Math. Proc. Cambridge Philos. Soc.* **114** (1993), 85–97.

[48] Hilsum, M., and Skandalis, G., Morphismes K-orientes d'espaces de feuilles et fonctorialité en théorie de Kasparov (d'aprés une conjecture d'A. Connes). *Ann. Sci. École Norm. Sup.* (4) **20** (3) (1987), 325–390.

[49] Hilsum, M., and Skandalis, G., Invariance par homotopie de la signature à coefficients dans un fibré presque plat. *J. Reine Angew. Math.* **423** (1992), 73–99.

[50] Johnson, W. B., and Randrianarivony, N. L., l_p ($p > 2$) does not coarsely embed into a Hilbert space. Preprint, 2004.

[51] Kasparov, G. G., The operator K-functor and extensions of C^*-algebras. *Izv. Akad. Nauk SSSR Ser. Mat.* **44** (3) (1980), 571–636; English transl. *Math. USSR-Izv.* **16** (1981), 513–572.

[52] Kasparov, G. G., Equivariant KK-theory and the Novikov conjecture. *Invent. Math.* **91** (1988), 147–201.

[53] Kasparov, G. G., and Skandalis, G., Groups acting properly on "bolic" spaces and the Novikov conjecture. *Ann. of Math.* (2) **158** (1) (2003), 165–206.

[54] Kasparov, G. G., and Yu, G., The coarse geometric Novikov conjecture and uniform convexity. *Adv. in Math.*, to appear.

[55] Kasparov, G. G., and Yu, G., Groups with coarse positive duality and the Novikov conjecture. Preprint, 2005.

[56] Lafforgue, V., K-théorie bivariante pour les algèbres de Banach et conjecture de Baum-Connes. *Invent. Math.* **149** (1) (2002), 1–95.

[57] Mendel, M., and Naor, A., Metric Cotype. Preprint, 2005.

[58] Mineyev, I., and Yu, G., The Baum-Connes conjecture for hyperbolic groups. *Invent. Math.* **149** (1) (2002), 97–122.

[59] Mishchenko, A. S., Infinite-dimensional representations of discrete groups, and higher signatures. *Izv. Akad. Nauk SSSR Ser. Mat.* **38** (1974), 81–106; English transl. *Math. USSR-Izv.* **8** (1974), 85–111.

[60] Mishchenko, A. S., and Fomenko, A. T., The index of elliptic operators over C^*-algebras. *Izv. Akad. Nauk SSSR Ser. Mat.* **43** (4) (1979), 831–859; English transl. *Math. USSR-Izv.* **15** (1) (1980), 87–112.

[61] Nowak, P., Coarsely embeddable spaces without property A. Preprint, 2005.

[62] Osin, D. V., Asymptotic dimension of relatively hyperbolic groups. Preprint, 2004.

[63] Ozawa, N., Amenable actions and exactness for discrete groups. *C. R. Acad. Sci. Paris Ser. I Math.* **330** (8) (2000), 691–695.

[64] Ozawa, N., A note on non-amenability of $B(l_p)$ for $p = 1, 2$. *Internat. J. Math.* **15** (6) (2004), 557–565.

[65] Ozawa, N., Boundary amenability of relatively hyperbolic groups. Preprint, 2005.

[66] Puschnigg, M., The Kadison-Kaplansky conjecture for word-hyperbolic groups. *Invent. Math.* **149** (1) (2002), 153–194.

[67] Roe, J., *Coarse cohomology and index theory for complete Riemannian manifolds,* Mem. Amer. Math. Soc. **104** (No. 407) (1993).

[68] Roe, J., *Index Theory, Coarse Geometry, and Topology of Manifolds.* CBMS Reg. Conf. Ser. Math. 90, Amer. Math. Soc., Providence, RI, 1996.

[69] Rosenberg, J., C^*-algebras, positive scalar curvature and the Novikov Conjecture, *Inst. Hautes Études Sci. Publ. Math.* **58** (1983), 197–212.

[70] Rosenberg, J., Analytic Novikov for topologists. In *Novikov conjectures, index theorems and rigidity* (Oberwolfach, 1993), Vol. 1, London Math. Soc. Lecture Note Ser. 226, Cambridge University Press, Cambridge 1995, 338–372.

[71] Sela, Z., Uniform embeddings of hyperbolic groups in Hilbert spaces. *Israel J. Math.* **80** (1–2) (1992), 171–181.

[72] Shan, L., An equivariant higher index theory and non-positively curved manifolds. Preprint, 2005.

[73] Stolz, S., Manifolds of positive scalar curvature. In *Topology of high-dimensional manifolds* (Trieste, 2001), No. 2, ICTP Lect. Notes 9, Abdus Salam Int. Cent. Theoret. Phys., Trieste 2002, 661–709.

[74] Skandalis, G., Tu, J.-L., and Yu, G., The coarse Baum-Connes conjecture and groupoids. *Topology* **41** (4) (2002), 807–834.

[75] Tu, J.-L., La conjecture de Baum-Connes pour les feuilletages moyennables. *K-Theory* **17** (3) (1999), 215–264.

[76] Tu, J.-L., Remarks on Yu's "property A" for discrete metric spaces and groups. *Bull. Soc. Math. France* **129** (1) (2001), 115–139.

[77] Weinberger, S., Aspects of the Novikov conjecture. In *Geometric and topological invariants of elliptic operators* (Brunswick, ME, 1988), Contemp. Math. 105, Amer. Math. Soc., Providence, RI, 1990, 281–297.

[78] Yu, G., Coarse Baum–Connes conjecture. *K-Theory* **9** (1995), 199–221.

[79] Yu, G., Localization algebras and the coarse Baum–Connes conjecture. *K-Theory* **11** (1997), 307–318.

[80] Yu, G., The Novikov Conjecture for groups with finite asymptotic dimension. *Ann. of Math.* (2) **147** (1998), 325–355.

[81] Yu, G., The coarse Baum-Connes conjecture for spaces which admit a uniform embedding into Hilbert space. *Invent. Math.* **139** (1) (2000), 201–240.

[82] Yu, G., Hyperbolic groups admit proper affine isometric actions on l^p-spaces. *Geom. Funct. Anal.* **15** (5) (2005), 1144–1151.

[83] Zuk, A., Property (T) and Kazhdan constants for discrete groups. *Geom. Funct. Anal.* **13** (3) (2003), 643–670.

Department of Mathematics, 1326 Stevenson Center, Vanderbilt University, Nashville, TN 37240, U.S.A.

E-mail: gyu@math.vanderbilt.edu

On spectral invariants in modern ergodic theory

Oleg N. Ageev *

Abstract. This is a short survey of recent developments in one of the oldest areas of ergodic theory, sometimes called the spectral theory of dynamical systems. We mainly discuss the spectral realization problem in the rich class of all invertible measure preserving dynamical systems, a "behavior" of different spectral invariants in natural subclasses of dynamical systems, and a complete solution of Rokhlin's problem on homogeneous spectrum in ergodic theory.

Mathematics Subject Classification (2000). Primary 37A, 28D; Secondary 47A05, 47A35, 47D03.

Keywords. Ergodic theory, spectral invariants, homogeneous spectrum problem.

1. Introduction

Ergodic theory, also called *measurable dynamics*, studies iterations of a map (a collection of maps forming an image of some (semi) group G with respect to a homomorphism) equipped with at least one (quasi) invariant measure. The maps are called *transformations, automorphisms*, or *dynamical systems* (the collection of maps is called a *G-action*, or a *dynamical system*, where G plays a role of "time").

One of the main classes of dynamical systems is the group of all invertible measure-preserving maps (automorphisms of a σ-algebra of measurable sets) defined on a non-atomic Lebesgue space (X, μ), $\mu(X) = 1$, where the Lebesgue space (X, μ) can be viewed as an interval $[0, 1]$ with the Lebesgue measure. This group is a Polish (complete metrizable separable) topological group, denoted by $\mathrm{Aut}(\mu)$, with respect to the weak (coarse) topology defined by

$$T_n \to T \Leftrightarrow \mu(T_n^{-1} A \Delta T^{-1} A) \to 0 \text{ for each measurable } A$$

(we identify transformations if they are coincide up to a measure zero set). We say that a property holds for a *typical* element from a topological space D if the set of elements with this property contains a dense G_δ-subset of D; in particular, for Polish spaces D a property holds for a *typical* element if the set of such elements is a comeager set. Iterations of $T \in \mathrm{Aut}(\mu)$ define a $\mathbb{Z}$-action denoted by the same

*The author is grateful to the Max Planck Institute (Bonn), the University of New South Wales (Australia), and the IHES (France) for hospitality during the writing and the correction of this paper.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

symbol T. More generally, given a discrete countable group G, the set of all G-actions, i.e. homomorphisms $T\colon G \to \mathrm{Aut}(\mu)$, equipped with the weak topology defined by

$$T(n) \to T \Leftrightarrow (\forall g \in G)\ T_g(n) \to T_g,$$

forms a Polish space denoted by Ω_G.

We are interested in *metrical* properties (equivalently, *metrical invariants*) of dynamical systems, i.e. in properties which are invariant with respect to measurable isomorphisms. Since every Polish space equipped with a Borel probability measure is measurably isomorphic to a Lebesgue space (X, μ) (not necessarily non-atomic, but usually we have a very easy dynamic on an atomic part), a "large" class of dynamical systems can be viewed as elements of $\mathrm{Aut}(\mu)$ or Ω_G.

If we wish to consider non-discrete groups G (the most famous example here is the case of $\mathbb{R}$, called a *flow*), then by G-action we mean only continuous homomorphisms (and their images) into $\mathrm{Aut}(\mu)$. As a rule it is no problem to find an invariant measure for a more or less natural map preserving some space. So dynamical systems are sufficiently common in different areas of mathematics. We are not able to discuss every result on this very rich theory. So we restrict ourselves to the main 'classical' subclass of one-to-one dynamical systems, i.e. to $\mathrm{Aut}(\mu)$ or Ω_G including, for example, shifts and automorphisms of commutative compact groups, interval exchanges, and billiards.

There exist many natural (or not so natural) ways to associate operators with dynamical systems. We will discuss the historically first classical operator introduced by Koopman [26]. A unitary operator $\hat{T}f(x) = f(Tx)$ induced on $L_2(X, \mu)$ by a transformation $T \in \mathrm{Aut}(\mu)$ is called *Koopman operator*. Unitary isomorphisms of Hilbert spaces define a spectral equivalence relation on the set of Koopman operators and, consequently, on $\mathrm{Aut}(\mu)$ or Ω_G.

Spectral invariants of dynamical systems, i.e. properties depending only on classes of the spectral equivalence relation, have become classical in the theory of dynamical systems after the celebrated work of John von Neumann [29] on classical mechanics.

Let us recall the properties

1. *to be ergodic* $\Leftrightarrow [\forall$ measurable $A[(\forall\, g \in G)[T_g A = A] \Rightarrow \mu(A)\mu(X\backslash A) = 0]]$,
2. *to be weakly mixing* $\Leftrightarrow [\exists g_n \to \infty[\hat{T}_{g_n} \to \int]]$,
3. *to be mixing* $\Leftrightarrow [\forall g_n \to \infty[\hat{T}_{g_n} \to \int]]$,

where we consider the weak operator topology and $\int$ is an operator of the orthogonal projection onto the space of constants. These are natural examples of (non-complete) spectral invariants, because each of them divides $\mathrm{Aut}(\mu)$ (resp. Ω_G) into only two subsets consisting of full classes of the spectral equivalence relation.

Consider a pair (ν, M), where ν is a Borel measure on $\mathbb{T}$ and M is a ν measurable function $M\colon \mathbb{T} \to \mathbb{N} \cup \{\infty\}$. Let $H^{\nu,M}$ be the subspace of the ν measurable square integrable functions $\varphi\colon \mathbb{T} \to l_2$ such that at a point $\lambda \in \mathbb{T}$ all but the first $M(\lambda)$

coordinates of $\varphi(\lambda)$ vanish. The space $H^{\nu,M}$ is a separable Hilbert space with respect to the inner product

$$\langle \varphi, \psi \rangle = \int_{\mathbb{T}} (\varphi(\lambda), \psi(\lambda))_{l_2} \, d\nu,$$

and the group $\mathbb{Z}$ acts unitarily on $H^{\nu,M}$ by the natural scalar multiplications:

$$U_n^{\nu,M} \varphi(\lambda) = \lambda^n \varphi(\lambda).$$

Due the spectral theorem for any unitary representation of $\mathbb{Z}$, say U, in a separable Hilbert space, there exists a pair (ν, M) such that $U^{\nu,M}$ is unitarily equivalent to U. This pair (ν, M) is unique in the sense that representations U^{ν_1,M_1} and U^{ν_2,M_2} are unitarily equivalent if and only if the measures ν_1 and ν_2 are equivalent and $M_1 = M_2$ almost everywhere with respect to ν_1 (or ν_2). Therefore a pair $([\nu], M(\hat{T}))$ is an example of a complete spectral invariant, where the *maximal spectral type* $[\nu]$ of $\hat{T}$ (or T) is a class of mutually equivalent Borel measures on $\mathbb{T}$ and $M(\hat{T})$ modulo ν-zero sets is the *spectral multiplicity function*. More concerning definitions of different spectral properties common in dynamics can be found in [13].

The setting of "*freeness*", where one says that a G-action is *free* if $\mu\{x \in X : (\exists g \neq 1)\ T_g x = x\} = 0$, can be viewed as the first trivial example of a non-spectral metrical invariant. However, usually in ergodic theory one considers only ergodic dynamical systems, because the "typical" problem in this area can be naturally reduced to the "ergodic" one. Since every ergodic transformation is obviously free, this setting is trivial in the class of ergodic $T \in \mathrm{Aut}(\mu)$. However there are many theorems going back to Anzai who showed that the spectral invariant is not a complete metrical invariant even in sufficiently "small" subclasses of $\mathrm{Aut}(\mu)$, and there are a few non-spectral metrical invariants. Let us especially remind here that the notion of *entropy* introduced by Kolmogorov [25] is one of those. In spite of these facts, in modern ergodic theory spectral invariants form a very rich class, and their study sometimes has essential influence on different branches of modern mathematics.

2. General classical problems of the spectral ergodic theory

1. *Spectral realization problem:* to determine precise conditions on the spectrum, i.e. spectral invariants, of a unitary operator under which it can be realized by a dynamical system.

2. *Spectral computation problem:* to calculate spectral invariants for every transformation of the Lebesgue space.

3. *Spectral isomorphism problem:* to describe the "gap" between spectral and metrical properties, i.e., to realize what kind of extra information should be added to spectral invariants for the purpose of having a complete metrical invariant.

There are numerous theorems, examples and counterexamples to different conjectures, which provides evidence of the extreme difficulty of handling these problems, because $\mathrm{Aut}(\mu)$ is a very rich class from the spectral point of view. This implies the importance of studying of general spectral problems in natural subclasses of dynamical systems, where such problems have not been solved yet.

3. Weak operator convergence

From time to time in the spectral ergodic theory was a pulsing activity stimulated by coming of new ideas. In the 60s and 70s some very importants problems were solved by approximations. We will concentrate our attention on $M(T)$ (i.e. $M(\hat{T})$). All known constructions of transformations T with calculations of nontrivial $M(T)$ are quite complicated. Oseledec [31] constructed a first example of an ergodic transformation T with $2 \le m(T) < 30$ where

$$m(T) = \sup\{n : n \in M(T)\}.$$

For a given number n, Robinson [33] used approximations for calculations and found an ergodic (even weakly mixing) transformation T with $m(T) = n$. More information about the history of the spectral multiplicity in ergodic theory can be found in [33], [27].

The main problem of the realization of possible values of the spectral multiplicity function was the homogeneous spectrum problem:

For any $n > 1$, does there exist an ergodic transformation T with $M(T) = \{n\}$ in the orthogonal complement of the space of constants?

This is a so-called Rokhlin problem on homogeneous spectrum. There is no written confirmation signed by V. Rokhlin of the fact that he suggested this problem. However in the memory of many (former) Russian mathematicians this long standing problem is associated with V. Rokhlin who posed this question in many discussions. Moreover, he even asked for less, namely

whether an ergodic transformation can have homogeneous spectrum with multiplicity different from 1 *or infinity in the orthogonal complement of the space of constants*.

The main goal of this section is to advertise a new sufficiently recent approach which uses weak operator convergence. This approach recently allowed the possibility to solve (Ageev, Ryzhikov) Rokhlin's old problem and to answer some other well-known questions, for example, Katok's question.

3.1. The case $1 \in M(T)$

Theorem 3.1. *For any $M \subseteq \mathbb{N} \cup \{\infty\}$ $(1 \in M)$ there exists an ergodic automorphism T with $M(T) = M$.*

Theorem 3.1 was proved in [27] for natural factors of transformations constructed in [19]. However the proof of the main result in [19] used, in fact, limit points (with respect to the weak operator convergence) of iterations of $\hat{T}$ in $\hat{T}$-invariant subspaces of $L_2(X, \mu)$. We remark here that Theorem 3.1 was also proved independently in [3] in the framework of a new direct construction.

Theorem 3.1 can be viewed as the complete solution of the realization problem of $M(T)$ in $L_2(X, \mu)$ for ergodic transformations T, because an easy corollary of the fact that the space of invariant functions is one dimensional is that $1 \in M(T)$.

3.2. Cartesian powers. For a typical T, Katok (see [21] and a renewed version [22]) showed using some approximation methods that

$$M(T \times T) \in \{\{2\}, \{2, 4\}\}$$

and conjectured that

$$M(\underbrace{T \times \cdots \times T}_{n \text{ times}}) = \{n, n(n-1), \ldots, n!\}.$$

It was the unique known construction, where a solution of Rokhlin's problem (only for $n = 2$) was expected.

This conjecture was confirmed completely by Ageev [2] (see also [6]) and, independently, by Ryzhikov [35] for $n = 2$, thereby giving a positive answer to Rokhlin's problem for $n = 2$. The solution of this version of Rokhlin's problem was a sufficiently simple application of the very unexpected fact that, for a typical transformation $T \in \mathrm{Aut}(\mu)$, all polynomials $P_n(\hat{T})$ forming a convex combination of some $\hat{T}^k$'s are limit points (with respect to the weak operator convergence) of iterations of $\hat{T}$.

This approach via so-called limit polynomials gave a lot of new additional information about $M(T)$. However, it did not give a possibility to solve the homogeneous spectrum problem for arbitrary n.

4. The homogeneous spectrum problem of arbitrary multiplicity

The homogeneous spectrum problem was solved completely in [7].

Theorem 4.1. *For any n there exists an ergodic transformation with homogeneous spectrum of multiplicity n in the orthogonal complement of the constant functions.*

One of the new guesses to prove this theorem was to apply the inner group symmetry coming from appropriate identities to symmetry in the spectrum of elements of corresponding group actions. Consider the following (quite exceptional for our purposes) group:

$$G_n = \mathrm{gr}\langle t, s; t_i t_j t_i^{-1} t_j^{-1} = 1 = t_0 \ldots t_{n-1} \text{ for all } i, j\rangle,$$

where $t_i = s^i t s^{-i}$, $t_0 = t$.

Theorem 4.1 is a natural corollary of the following main theorem:

Theorem 4.2 ([7]). *For a typical G_n-action T, T_{s^n} is a weakly mixing transformation with homogeneous spectrum of multiplicity n in the orthogonal complement of the constant functions.*

4.1. Sketch of Proof. By our choice G_n contains a normal subgroup G of finite index so that G is a free commutative group in generators $t_0, \dots, t_{n-2}, s^n$. The spectral theorem applied for G-subaction of a G_n-action T implies that the Hilbert space $L_2(X, \mu)$ can be decomposed into the direct sum of mutually orthogonal T_{s^n}-invariant subspaces $H_i, i = 1, \dots, n$, and a remaining part L, so that both restrictions of T_{s^n} to H_i are mutually unitarily equivalent and L is a subspace of $L_2(X, \mu)$ spanned by all eigenfunctions of T_g, $g \in G$. Therefore if L is trivial, i.e. it consists of constant functions, then the values of the spectral multiplicity function of T_{s^n} are multiples of n in the orthogonal complement of the constant functions.

By our choice G_n is sufficiently "close" to the class of commutative groups. This implies the guess that, given $g \in G_n$, the topological status in Ω_{G_n} of different spectral invariants of T_g is "almost" the same as in $\mathrm{Aut}(\mu)$. More precisely, we found counterparts in Ω_{G_n} of well-known classical results that, for a typical T in $\mathrm{Aut}(\mu)$, the transformations T^k ($k \neq 0$) are weakly mixing and have a simple spectrum. Namely, for a typical G_n-action we prove two statements: (I) Transformations T_s have a simple spectrum. (II) T_g ($g \neq 1$)[1] are weakly mixing (or, equivalently, have no eigenfunctions in the orthogonal complement of the constant functions).

To show (II) we follow a classical method to prove that some property A is typical in the space of all actions of some group. Namely, we find a free action having the property A. Rokhlin's lemma usually implies that the set of conjugations of any fixed free action is dense. Since we only consider properties which are invariant with respect to any metrical isomorphism (in particular, any conjugation), the set of actions having the property A is dense. Finally, we use convenient approximations by elements of this dense set to show that A is valid for a typical action. This method explains the importance to construct examples of dynamical systems having different metrical properties in ergodic theory.

A priori, we have no example of a free G_n-action T ($n \geq 3$) such that T_s has a simple spectrum. Therefore to show (I) we applied another (more delicate) method that used cyclic approximations [24][2]. Namely, we proved that a typical G_n-action T can be approximated by appropriate finite G_n-actions $T(k)$ so that T_s is approximated sufficiently fast by a cyclic $T_s(k)$.

[1] Formally, (II) was proved in [7] only for a certain subset of elements of G. This was sufficient to establish Theorem 4.2. However, following the same method of the proof, we can then deduce almost automatically that (II) holds.

[2] A very interesting recent preprint of Danilenko contains a few technical improvements. Using our method, he proved Theorem 4.2 for a simplified version G^* of G, and, in particular, he replaced arguments of cyclic approximations by algebraic ones. He also constructed explicit examples of ergodic transformations with homogeneous spectrum of multiplicity n on L_2^0 using (not directly) the well-known fact that transformations admitting a sufficiently fast approximation can be constructed explicitly. Proofs for G^* are simpler, but my choice of G_n in [7] was made due to different reasons, for example, to show that even in the well-studied class of transformations conjugated to their inverses (i.e. $T_t, t \in G_2$) we have a few new interesting results.

Finally, to have a homogeneous spectrum of multiplicity n, we note that L is trivial, because there is no non-constant eigenfunction for any T_g ($g \neq 1$) if the G_n-action T is typical, and the upper bound n of $M(T_{s^n})$ follows from simplicity of the spectrum of T_s.

I would like to stress that an analogy between Ω_{G_n} and $\mathrm{Aut}(\mu)$ is not always certain, because, for example, for a typical G_n-action T, T_s^n does not have a simple spectrum.

4.2. Some spectral properties of typical G_n-actions. Using a technique from [7], we can say somewhat more concerning elements of a typical G_n-action. More precisely, the first simple observation in the proof of the main theorem in [7] is the remark that for a typical T from Ω_{G_n}, T_s (equivalently, T_s^n) is a *rigid* (non-mixing) transformation, i.e. $T_t^{k_i} \to E$ for some $k_i \to \infty$ (E the identity map). In particular, this means that a σ_{T_s} (or $\sigma_{T_{s^n}}$) is singular, where σ_T is the measure of the maximal spectral type of T.

It is easy to check that for the transformations in Theorem 4.2, $\sigma'_{T_s} \sim \varphi\sigma'_{T_s}$, where $\varphi\colon \mathbb{T} \to \mathbb{T}$ is a map defined by $\varphi(\lambda) = \exp(2\pi i/n)\lambda$, and σ'_S is the measure of the maximal spectral type of $\hat{S}$ on $\{\mathrm{const}\}^\perp$. Therefore we have the following corollary:

Corollary 4.3 ([7]). *For a typical G_n-action T, if $k|n$ then T_s^k is a weakly mixing transformation with homogeneous spectrum of multiplicity k in the orthogonal complement of the constant functions.*

Analogous results for an element T_t are somewhat surprising.

Theorem 4.4 ([7]). *For a typical T from Ω_{G_n}, the following properties hold:*

1. *T_t is weakly mixing.*
2. *T_t is rigid.*
3. *If $n = 2$, then T_t has a homogeneous spectrum of multiplicity 2 on $\{const\}^\perp$.*
4. *If $n > 2$, then T_t has a simple spectrum.*

5. Spectral rigidity of group actions

Our discussion in this section focuses on values of the spectral multiplicity function of elements of typical dynamical systems.

It is easy to see that $M(T_g) = \{\infty\}$ if g has a finite order.

Remark 5.1. Let G be a countable abelian group. Then, for a typical G-action T, $M(T_g) = \{1\}$ if g has infinite order.

The proof is more or less a standard application of approximation arguments.

All discussion above suggests us to conclude that, as a rule, groups fulfill some (new) effect which I call spectral rigidity.

Definition 5.2. Following [7] we say that an element h of a group H has *spectral rigidity* if for a typical H-action T the set of essential values $M(\hat{T}_h)$ of the spectral multiplicity function of T_h is constant. If every element of H has this property we say that the group H has *spectral rigidity*.

It is easy to see that the notion of spectral rigidity can be considered as an invariant with respect to group or metrical isomorphisms.

In fact, the main theorem on the homogeneous spectrum could be proved because certain elements of G_n have spectral rigidity. Let us mention a partial result valid for every countable group.

Proposition 5.3. *Suppose g is an element of a countable group G; then $m(T_g)$ is independent of our choice of T from Ω_G except for a meager set of actions.*[3]

This proposition is an easy corollary of Proposition 3 in [7] and the fact that every countable group G has the *weak Rokhlin property*, i.e. $\{T \in \Omega_G : T \cong T'\}$ is dense for some $T' \in \Omega_G$, where $\cong$ is a metrical isomorphism established by Glasner, Thouvenot, and Weiss in [16].

Let us also mention that an easy application of the technique to study spectral rigidity [7] to solvable Baumslag–Solitar groups is the following theorem (cf. [9]).

Theorem 5.4. *For a typical G-action T, T_t is a weakly mixing rank one transformation, where $G = \mathrm{gr}\langle t, s;\ ts = st^2\rangle$.*

This theorem gives a positive answer to a well-known question (see, for example, [18]).

6. Spectral invariants in natural subclasses of dynamical systems

6.1. Finite rank case. Let us recall one of the metrical invariants introduced by Ornstein and Chacon and intensively studied over the last three decades.

Definition 6.1. The transformation T has *rank n* if n is the smallest number such that, for any k, there exist towers (columns)

$$A_{k,i}, TA_{k,i}, \ldots, T^{h_{k,i}}A_{k,i}, \quad i = 1, \ldots, n$$

such that all levels $T^j A_{k,i}$ and the remaining set form a measurable partition ξ_k of X and $\xi_k \to \varepsilon$, i.e., for any measurable set A there are ξ_k-measurable sets A_k such

[3] Recently, this proposition was extended in [10] to the case $M(T_g)$, i.e. every countable group has spectral rigidity. However, the calculation of values of $M(T_g)$ for a typical group action T remains an interesting unsolved problem even in the class of countable groups.

that $\mu(A\Delta A_k) \to 0$ as $k \to \infty$. A transformation has *infinite* rank if there is no such number. Finally, a transformation has *uniform* rank n if the towers above can be chosen in such a way that $h_{k,i}$ is independent of i.

Let us mention a few results on spectral invariants in the case of finite rank transformations. Chacon [14] proved that this class does not contain transformations with an unbounded spectral multiplicity function since always

$$m(T) \leq \text{rank } T.$$

Ornstein [30] proved that there exist rank one mixing transformations (see also another staircase construction of mixing rank one transformations [1]). Bourgain [12] showed that Ornstein's transformations have a singular spectrum. In [4] it was proved that there exist finite rank mildly mixing transformations with a Lebesgue component of any even multiplicity in the spectrum (see also [32],[28]). Every finite subset of $\mathbb{N}$ (with 1) can be a spectral multiplicity function of finite rank transformations (see [3]).

Observe that the setting to be of finite rank can be viewed as the easiest geometrical version of an appropriate approximation. Therefore the solution of the homogeneous spectrum problem in the class of finite rank transformations was a natural addition to the main result of [7]. More precisely, we have the following result.

Theorem 6.2 ([7]). *For a typical G_n-action T, T_{s^n} is a weakly mixing transformation with homogeneous spectrum of the multiplicity n in the orthogonal complement of the constant functions. Moreover, T_{s^n} has uniform rank n.*

Remark 6.3. All transformations of the form $T \times T$ have infinite rank (Ryzhikov [36]). Therefore the transformations in Theorem 4.2 (and 4.4) are from a new class with homogeneous spectra.

6.2. Mixing case In spite of the fact that in the set of all ergodic transformations we have almost full information concerning possible values of the spectral multiplicity function, the case of mixing transformations is still weakly studied, because many useful methods working for typical transformations are not applicable to the subset of mixing transformations.

The complete calculation of values of the spectral multiplicity function of natural factors of the Cartesian power of a transformation (under certain conditions on a transformation) was given in [5]. This construction can be considered as a source of mixing transformations having new (highly nonhomogeneous) spectral multiplicity functions, and covers all currently known non-trivial (i.e. $M(\hat{T}) \neq \{1\}, \{\infty\}$) examples of $M(\hat{T})$ for mixing transformations.

Let G be any subgroup of the *symmetric* group $\mathfrak{S}_n$ acting on $\{1, \ldots, n\}$ by permutations. For any $1 \leq k \leq n$ denote by the same symbol G the *diagonal* action of G on $I_k = \{1, \ldots, n\}^k$. Consider a restriction of the G-action to a G-invariant subset

$$I_k' = \{i_k = (i_k(1), \ldots, i_k(k)) \in I_k : i_k(l) = i_k(m) \text{ iff } l = m\}.$$

Suppose $\sim$ is an orbital equivalence relation naturally defined by G on I'_k. Define

$$D_k = \sharp I'_k/\sim\ (1 \le k \le n) \quad \text{and} \quad M_G(n) = \{D_1, \dots, D_n\}.$$

Theorem 6.4 ([5]). *For any subgroup G of $\mathfrak{S}_n$ there exists a mixing (of all orders) transformation T satisfying*

$$M(\hat{T}|_{L_2^0}) = M_G(n),$$

where $L_2^0 = \{f \in L_2 : \int f\,d\mu = 0\}$.

Corollary 6.5. *For any n there exists a mixing (of all orders) transformation T satisfying*

$$M(\hat{T}) = \{2, 3, \dots, n\} \quad \text{on } L_2^0.$$

This gives

$$M(\hat{T}) = \{1, 2, 3, \dots, n\} \quad \text{on } L_2.$$

Let us remark that for the case $n = 2$ this was proved earlier by Ryzhikov (see [37]), and that Corollary 6.5 also gives an answer to a question of Robinson (see [34] and [17], 5.3) about the set of possible values of $m(T) = \max M(\hat{T})$ for mixing transformations.

6.3. Interval exchanges Let us remind the definition of interval exchange (transformation). Let $n > 1$ and let π be an irreducible permutation of $\{1, \dots, n\}$, where a permutation π is called *irreducible* if $\pi\{1, \dots, m\} \ne \{1, \dots, m\}$, $1 \le m < n$. Let Δ be the simplex in $\mathbb{R}^n$,

$$\lambda = (\lambda_1, \dots, \lambda_n), \quad \lambda_i \ge 0,\ \sum_i \lambda_i = 1.$$

The unit interval $I = [0, 1)$ is divided into semi-open intervals

$$I_m = \Big[\sum_{i<m} \lambda_i, \sum_{i\le m} \lambda_i\Big), \quad 1 \le m \le n.$$

Definition 6.6. The *interval exchange transformation* $T_{\pi,\lambda}$ is uniquely defined on every I_m as a *shift*, i.e. a map $T_m\colon x \to x + \alpha_m$ such that that the intervals I_m are rearranged according to the permutation π, where α_m is

$$\sum_{i:\pi\{i\}\le\pi\{m\}} \lambda_i - \sum_{i\le m} \lambda_i.$$

If π is a *rotation*, i.e. $\pi(i+1) \equiv \pi(i) + 1 \pmod n$, then $T_{\pi,\lambda}$ has a very easy dynamical behavior, because it is a shift on $I \equiv \mathbb{R}/\mathbb{Z}$. Given an irreducible permutation π which is not a rotation, we restrict ourselves to the study of the topological status of subsets of transformations with fixed spectral invariants. Then the class of all interval

exchanges $T_{\pi,\lambda}$, say Ω_π, looks like a small model of $\mathrm{Aut}(\mu)$. However, due a natural parametrization of Ω_π by elements of Δ, it is more natural to consider the setting "to be typical" from the measure theoretical point view, i.e. to say that a property A is "typical" in Ω_π if the set of $\lambda \in \Delta$ such that $T_{\pi,\lambda}$ satisfies A forms a set of full Lebesgue measure. Indeed, analogs of the very well-known results that a typical transformation $\mathrm{Aut}(\mu)$ is not mixing, weakly mixing, and has a simple spectrum are the following. Katok [23] showed that there do not exist mixing interval exchanges. Veech [38] proved that a "typical" interval exchange has a simple spectrum. And only recently Avila and Forni [11] solved the long-standing problem that a "typical" interval exchange is weakly mixing.

As an analog to Theorem 3.1 it was proved in [8] that for any bounded M, $1 \in M \subseteq \mathbb{N}$, there exists an ergodic interval exchange T with $M(T) = M$ in the class of all interval exchanges. Using Oseledec's [31] result that the upper bound of the spectral multiplicity of an ergodic interval exchange of n intervals is at most n, we see that this is a complete solution of the problem of possible values of the spectral multiplicity function ($1 \in M(T)$) in the subclass of ergodic interval exchanges.

No ergodic interval exchange T is known with $1 \notin M(T)$.

7. Some more problems

Spectral invariants were studied in ergodic theory for more than seven decades and a lot of unsolved problems have been accumulated. Let me formulate some of them.

7.1. (One of the oldest problems in ergodic theory associated with Banach.) Does there exist a (mixing) transformation with simple Lebesgue spectrum in the orthogonal complement of constant functions?

7.2. (A well-known problem.) Is it true that every rank one transformation has a singular spectrum?

7.3. Is the property to be mixing of all orders spectral (i.e. a spectral invariant)?

7.4. (Another well-known problem.) Is the property to have a fixed rank spectral?

7.5. Given $n > 1$, find the set of pairs (g, G), where G is a countable group and $g \in G$ such that, for a typical G-action T, a transformation T_g has a homogeneous spectrum of multiplicity n in the orthogonal complement of the constant functions.

References

[1] Adams, T., Smorodinsky's conjecture on rank-one mixing. *Proc. Amer. Math. Soc.* **126** (1998), 739–744.

[2] Ageev, O. N., On ergodic transformations with homogeneous spectrum. *J. Dynam. Control Systems* **5** (1999), 149–152.

[3] Ageev, O. N., On the multiplicity function of generic group extensions with continuous spectrum. *Ergodic Theory Dynam. Systems* **21** (2001), 321–338.

[4] Ageev, O. N., Dynamical systems with a Lebesgue component of even multiplicity in the spectrum. *Math. Sb.* **64** (1989), 305–317.

[5] Ageev, O. N., Mixing with staircase multiplicity function. Preprint; MPIM2005-48.

[6] Ageev, O. N., On the spectrum of Cartesian powers of classical automorphisms. *Mat. Zametki* **68** (2000), 643–647; English transl. *Math. Notes* **68** (2000), 547–551.

[7] Ageev, O., The homogeneous spectrum problem in ergodic theory. *Invent. Math.* **160** (2005), 417–446.

[8] Ageev, O. N., The spectral multiplicity function and geometric representations of interval exchange transformations. (Russian) *Mat. Sb.* **190** (1999), 3–28; English transl. *Sb. Math.* **190** (1999), 1–28.

[9] Ageev, O. N., Spectral rigidity of group actions: Applications to the case $\mathrm{gr}\langle t, s; ts = st^2\rangle$. *Proc. Amer. Math. Soc.* **134** (2006), 1331–1338.

[10] Ageev, O., Spectral rigidity of group actions and Kazhdan's groups. Preprint.

[11] Avila, A., Forni, G., Weak mixing for interval exchange transformations and translations flows. *Ann. of Math.*, to appear.

[12] Bourgain, J., On the spectral type of Ornstein's class one transformations. *Israel J. Math.* **84** (1993), 53–63.

[13] Cornfel'd, I. P., Fomin, S. V., Sinai, Ya. G., *Ergodic Theory*. Grundlehren Math. Wiss. 245, Springer-Verlag, New York 1982.

[14] Chacon, R. V., Approximation and spectral multiplicity. In *Contributions to Ergodic Theory and Probability* (Columbus, Ohio, 1970), Lecture Notes in Math. 160, Springer-Verlag, Berlin 1970, 18–27.

[15] Danilenko, A. I., Explicit solution of Rokhlin's problem on homogeneous spectrum and applications. Preprint.

[16] Glasner, E., Thouvenot, J.-P., Weiss, B., Every countable group has the weak Rokhlin Property. *Bull. London Math.Soc.*, to appear.

[17] Goodson, G. R., A survey of recent results in the spectral theory of ergodic dynamical systems. *J. Dynam. Control Systems* **5** (1999), 173–226.

[18] Goodson, G. R., Ergodic dynamical systems conjugate to their composition squares. *Acta Math. Univ. Comenian.* (*N.S.*) **71** (2002), 201–210.

[19] Goodson, G. R., Kwiatkowski, J., Lemańczyk, M., and Liardet, P., On the multiplicity function of ergodic group extensions of rotations. *Studia Math.* **102** (1992), 157–174.

[20] Halmos, P. R., *Lectures on ergodic theory*. Chelsea Publishing Co., New York 1960.

[21] Katok, A. B., *Constructions in ergodic theory.* Unpublished lecture notes.

[22] Katok, A. B., *Combinatorial constructions in ergodic theory and dynamics*. Univ. Lecture Ser. 30, Amer. Math. Soc., Providence, RI, 2003.

[23] Katok, A. B., Interval exchange transformations and some special flows are not mixing. *Israel J. Math.* **35** (1980), 301–310.

[24] Katok, A. B., Stepin, A. M., Approximations in ergodic theory. *Russian Math. Surveys*. **22** (1967), 77–102.

[25] Kolmogorov, A. N., A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk SSSR* (*N.S.*) **119** (1958), 861–864 (in Russian).

[26] Koopman, B. O., Hamiltonian systems and transformations in Hilbert space, *Proc. Nat. Acad. Sci. U.S.A.*. **17** (1931), 315–318.

[27] Kwiatkowski, J., Lemańczyk, M., On the multiplicity function of ergodic group extensions. II. *Studia Math.* **116** (1995), 207–214.

[28] Mathew, J., Nadkarni, M. G., A measure preserving transformation whose spectrum has Lebesgue component of multiplicity two. *Bull. London Math. Soc.* **16** (1984), 402–406.

[29] Neumann, J. von, Zur Operatorenmethode in der klassischen Mechanik. *Ann. of Math.* **33** (1932), 587–642.

[30] Ornstein, D. S., On the root problem in ergodic theory. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley, Calif., 1970/1971), Vol. II: Probability theory, University of California Press, Berkeley, Calif., 1972, 347–356.

[31] Oseledec, V. I., The spectrum of ergodic automorphisms. *Dokl. Akad. Nauk SSSR* **168** (1966), 1009–1011 (in Russian).

[32] Queffelec, M., *Substitution dynamical systems—spectral analysis*. Lecture Notes in Math. 1294, Springer-Verlag, Berlin 1987.

[33] Robinson, E., Jr., Ergodic measure preserving transformations with arbitrary finite spectral multiplicity. *Invent. Math.* **72** (1983), 299–314.

[34] Robinson, E., Jr., Mixing and spectral multiplicity, *Ergodic Theory Dynam. Systems* **5** (1985), 617–624.

[35] Ryzhikov, V. V., Transformations having homogeneous spectra, *J. Dynam. Control Systems* **5** (1999), 145–148.

[36] Ryzhikov, V. V., On the ranks of an ergodic automorphism $T \times T$. *Funktsional. Anal. i Prilozhen.* **35** (2001), 84–87; English transl. *Funct. Anal. Appl.* **35** (2001), 151–153.

[37] Ryzhikov, V. V., Homogeneous spectrum, disjointness of convolutions, and mixing properties of dynamical systems. *Selected Russian Mathematics* **1** (1999), 13–24.

[38] Veech, W., The metric theory of interval exchange transformations. I., II., III., *Amer. J. Math.* **106** (1984), 1331–1359, 1361–1387, 1389–1422.

N.E. Bauman State Technical University, 105005, Moscow, Russian Federation
E-mail: ageev@mx.bmstu.ru

Ergodic Ramsey theory: a dynamical approach to static theorems

Vitaly Bergelson*

Abstract. We discuss classical results of Ramsey theory together with their dynamical counterparts, survey recent developments and formulate some natural open questions and conjectures.

Mathematics Subject Classification (2000). Primary 28D15, 05D10; Secondary 37A45, 37B10.

Keywords. Amenable group, ergodic theory, Furstenberg's correspondence principle, nilmanifold, Ramsey theory, recurrence, Stone–Čech compactification.

Introduction

Since its inception, ergodic theory was successfully employing combinatorial ideas and methods – from the use of the ubiquitous pigeonhole principle in the proof of the Poincaré recurrence theorem to combinatorial proofs of maximal inequalities, to the utilization of the marriage lemma in Ornstein's isomorphism theory. This debt to combinatorics is amply repaid by the accomplishments of the ergodic Ramsey theory.

Ergodic Ramsey theory was initiated in 1977 when H. Furstenberg [F3] proved a far reaching extension of the classical Poincaré recurrence theorem and derived from it the celebrated Szemerédi's theorem [Sz], which states that every set $E \subset \mathbb{N}$ with $\bar{d}(E) := \limsup_{N\to\infty} \frac{|E\cap\{1,2,\dots,N\}|}{N} > 0$ contains arbitrarily long arithmetic progressions. Furstenberg's ergodic approach to Szemerédi's theorem has not only revealed the dynamical underpinnings of this seemingly static result, but has also opened new vistas for mutually perpetuating research in ergodic theory, combinatorics, and number theory.

This survey is organized as follows. In Section 1 we formulate some of the classical theorems of Ramsey theory and discuss their dynamical counterparts. The subsequent sections are devoted to more recent developments and contain formulations of some natural open questions and conjectures. Unfortunately, due to space constraints, some of the topics will not get the attention they deserve. The readers will find more material together with much more elaborated discussion in the recent survey [B4]. (See also [F4], [B1] and [B3].)

*This work was supported in part by NSF grant DMS-0345350.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

1. A brief survey

In *partition Ramsey theory* the focus is on patterns which can always be found in one cell of any finite partition of a highly organized structure such as $\mathbb{Z}^d$, a complete graph, a vector space, etc.[1] Here are some examples.

Theorem 1.1 (Gallai–Grünwald[2]). *For all $r, d \in \mathbb{N}$, if $\mathbb{Z}^d = \bigcup_{i=1}^r C_i$, then one of C_i has the property that for every finite set $B \subset \mathbb{Z}^d$, there exist $n \in \mathbb{N}$ and $v \in \mathbb{Z}^d$ such that $v + nB = \{v + nu : u \in B\} \subset C_i$. In other words, C_i contains a homothetic image of every finite set.*

Theorem 1.2 (cf. [GraRS], p. 40). *Let V be an infinite vector space over a finite field. For all $r \in \mathbb{N}$, if $V = \bigcup_{i=1}^r C_i$, then one of C_i contains affine vector spaces of arbitrary finite dimension.*

Theorem 1.3 (Hindman's Theorem, [Hi]). *Given an infinite set $E = \{x_1, x_2, \dots\}$ of natural numbers let $FS(E) = \{x_{i_1} + x_{i_2} + \cdots + x_{i_k} : i_1 < i_2 < \cdots < i_k,\ k \in \mathbb{N}\}$. For all $r \in \mathbb{N}$, if $\mathbb{N} = \bigcup_{i=1}^r C_i$, then one of C_i contains a set of the form $FS(E)$ for some infinite set $E \subset \mathbb{N}$.*

The *density Ramsey theory* attempts to explain (and enhance) the results of the partition Ramsey theory by studying the patterns which ought to appear in any "large" set. The notion of largeness may vary but it is always assumed to be partition regular in the sense that for any finite partition of a large set at least one of the cells is large. One also assumes (of course) that the family of large subsets of the ambient structure S includes S itself. We will formulate now density results which correspond to (and refine) the partition results contained in Theorems 1.1, 1.2, and 1.3.

Theorem 1.4 (Furstenberg–Katznelson's multidimensional Szemerédi theorem, [FK1]). *Let $d \in \mathbb{N}$ and assume that $E \subset \mathbb{Z}^d$ is a set of positive upper density, that is*

$$\bar{d}(E) = \limsup_{N\to\infty} \frac{|E \cap [-N, N]^d|}{(2N+1)^d} > 0.$$

For every finite set $B \subset \mathbb{Z}^d$ there exist $n \in \mathbb{N}$ and $v \in \mathbb{Z}^d$ such that $v + nB \subset E$.

Let F be a finite field and let V_F be a countably infinite vector space over F. To define a notion of largeness which will allow us to formulate a density version of Theorem 1.2, observe that, as an abelian group, V_F is isomorphic to the direct sum F_∞ of countably many copies of F:

$$F_\infty = \{(a_1, a_2, \dots) : a_i \in F \text{ and all but finitely many } a_i = 0\} = \bigcup_{n=1}^{\infty} F_n,$$

[1] More precisely, one either deals with arbitrary finite partitions of an infinite structure or with partitions into a fixed number of cells of sufficiently large finite structures.

[2] Gruńwald (who later changed his name to Gallai) apparently never published his proof. See [R], p. 123 and [GraRS], p. 38.

where $F_n = \{(a_1, a_2, \dots) : a_i = 0 \text{ for } i > n\} \cong F \oplus \cdots \oplus F (n \text{ times})$. For a set $E \subset V_F$ (where V_F is identified with F_∞), define the upper density $\bar{d}(E)$ as[3]

$$\bar{d}(E) = \limsup_{n\to\infty} \frac{|E \cap F_n|}{|F_n|}.$$

Theorem 1.5. *Every set of positive upper density in the vector space V_F contains affine subspaces of every finite dimension.*

Before introducing a family of notions of largeness which is pertinent to Hindman's theorem (Theorem 1.3 above), let us remark that the naive attempt to use the notion of upper density (which works well for the Szemerédi-type theorems) immediately fails. Indeed, while the set of odd integers has density $\frac{1}{2}$, it clearly does not contain the sum of any two of its elements. One can actually construct, for any $\varepsilon > 0$, a set $S \subset \mathbb{N}$ with $d(S) = \lim_{N\to\infty} \frac{|S\cap\{1,2,\dots,N\}|}{N} > 1 - \varepsilon$, which does not contain $FS(E)$ for any infinite set $E \subset \mathbb{N}$.

It turns out that a natural notion of largeness appropriate for Hindman's theorem can be introduced with the help of $\beta\mathbb{N}$, the Stone–Čech compactification of $\mathbb{N}$. In view of the increasingly important role which Stone–Čech compactifications play in ergodic Ramsey theory, we will briefly discuss some of the relevant definitions and facts. For missing details see ([B1], Section 3) and [B3]. (See also [HiS] for a comprehensive treatment of topological algebra in Stone–Čech compactifications.)

A convenient way of dealing with $\beta\mathbb{N}$ is to view it as an appropriately topologized set of ultrafilters on $\mathbb{N}$. Recall that an *ultrafilter* p on $\mathbb{N}$ is a maximal filter, namely a family of subsets of $\mathbb{N}$ satisfying the following conditions (the first three of which constitute, for a nonempty family of sets, the definition of a *filter*).

(i) $\emptyset \notin p$;

(ii) $A \in p$ and $A \subset B$ imply $B \in p$;

(iii) $A \in p$ and $B \in p$ imply $A \cap B \in p$;

(iv) (maximality) if $r \in \mathbb{N}$ and $\mathbb{N} = \bigcup_{i=1}^{r} C_i$ then for some $i \in \{1, 2, \dots, r\}$, $C_i \in p$.

One can naturally identify each ultrafilter p with a finitely additive $\{0, 1\}$-valued probability measure μ_p on the power set $\mathcal{P}(\mathbb{N})$. This measure μ_p is defined by the requirement $\mu_p(C) = 1$ iff $C \in P$. Without saying so explicitly, we will always think of ultrafilters as such measures, but will prefer to write $C \in p$ instead of $\mu_p(C) = 1$.

Any $n \in \mathbb{N}$ defines the so-called *principal* ultrafilter $\{C \subset \mathbb{N} : n \in C\}$. Principal ultrafilters can be viewed as point measures corresponding to elements of $\mathbb{N}$, and are the only ones which can be constructed without the use of Zorn's lemma (see [CN], pp. 161–162). Since ultrafilters are maximal filters, any family of subsets of $\mathbb{N}$ which

[3]This definition depends, of course, on the way V_F is represented as an infinite direct sum. Each such representation leads to a notion of upper density in V_F.

has the finite intersection property can be "extended" to an ultrafilter. Given $C \subset \mathbb{N}$, let $\overline{C} = \{p \in \beta\mathbb{N} : C \in p\}$. The family $\mathcal{G} = \{\overline{C} : C \in \mathbb{N}\}$ forms a basis for the open sets of a topology on $\beta\mathbb{N}$ (as well as a basis for the closed sets), and, with this topology, $\beta\mathbb{N}$ is a compact Hausdorff space. Clearly, $\overline{\mathbb{N}} = \beta\mathbb{N}$, which hints that the operation of addition (and that of multiplication, as well) can be extended from $\mathbb{N}$ to $\beta\mathbb{N}$. In the following definition, $C - n$ (where $C \subset \mathbb{N}$ and $n \in \mathbb{N}$) is the set of all m such that $m + n \in C$. For $p, q \in \beta\mathbb{N}$, define

$$p + q = \{A \subset \mathbb{N} : \{n \in \mathbb{N} : (A - n) \in p\} \in q\} \tag{1}$$

It is not hard to check that for principal ultrafilters the operation $+$ corresponds to addition in $\mathbb{N}$. One can show that $p + q$ is an ultrafilter, that the operation $+$ is associative and that, for any fixed $p \in \beta\mathbb{N}$, the function $\lambda_p(q) = p + q$ is a continuous self-map of $\beta\mathbb{N}$. It follows that with the operation $+$, $\beta\mathbb{N}$ becomes a compact *left topological* semigroup. By a theorem due to R. Ellis, [E], any such semigroup has an idempotent. It turns out the idempotent ultrafilters in $(\beta\mathbb{N}, +)$ (viewed as measures) have a natural shift-invariant property which is responsible for a variety of applications including the following result which may be regarded as a density version of Hindman's theorem.

Theorem 1.6. *Let p be an idempotent ultrafilter in $(\beta\mathbb{N}, +)$. If $C \in p$, then there is an infinite set $E \subset \mathbb{N}$ such that $FS(E) \subset C$.*

While Theorems 1.1 through 1.6 obviously have a common Ramsey-theoretical thread, their formulations do not reveal much about their dynamical content. Our next goal is to convince the reader that all of these results can be interpreted as recurrence theorems in either topological or measure-preserving dynamics. (Topological dynamics forms the natural framework for partition results; measure preserving dynamics corresponds to density statements).

We start with formulating the dynamical version of the Gallai–Grünwald theorem. The idea to apply the methods of topological dynamics to partition results is due to H. Furstenberg and B. Weiss (See [FW]).

Theorem 1.7 (cf. [FW], Theorem 1.4). *Let $d \in \mathbb{N}, \varepsilon > 0$, and let X be a compact metric space. For any finite set of commuting homeomorphisms $T_i : X \to X, i = 1, 2, \ldots, k$, there exist $x \in X$ and $n \in \mathbb{N}$ such that* $\operatorname{diam}\{x, T_1^n x, T_2^n x, \ldots, T_k^n x\} < \varepsilon$.

To derive Theorem 1.1 from Theorem 1.7, one utilizes the fact that, for fixed $r, d \in \mathbb{N}$, the r-colorings of $\mathbb{Z}^d$ (viewed as mappings from $\mathbb{Z}^d$ to $\{1, 2, \ldots, r\}$) are naturally identified with the points of the compact product space $\Omega = \{1, 2, \ldots, r\}^{\mathbb{Z}^d}$. For $m = (m_1, m_2, \ldots, m_d) \in \mathbb{Z}^d$, let $|m| = \max_{1 \le i \le d} |m_i|$. Introduce a metric on Ω by defining, for any pair $x, y \in \Omega$, $\rho(x, y) = \inf_{n \in \mathbb{N}} \left\{\frac{1}{n} : x(m) = y(m) \text{ for } |m| < n\right\}$. Note that $\rho(x, y) < 1 \iff x(0) = y(0)$. Let $B = \{b_1, \ldots, b_k\} \subset \mathbb{Z}^d$. Define the homeomorphisms $T_i : \Omega \to \Omega$, $i = 1, 2, \ldots, k$ by $(T_i x)(m) = x(m + a_i)$, and set, for $l = (l_1, l_2, \ldots, l_k) \in \mathbb{Z}^k$, $T^l = T_1^{l_1} T_2^{l_2} \cdots T_k^{l_k}$. Let $y(m)$, $m \in \mathbb{Z}^d$, be the element

of Ω which corresponds to the given coloring $\mathbb{Z}^d = \bigcup_{i=1}^r C_i$. let $X = \overline{\{T^l y\}}_{l\in\mathbb{Z}^k}$ be the orbital closure of x in Ω. It follows from Theorem 1.7 that for some $x \in X$ which can without loss of generality be chosen to be of the form $T^u y$ for some $u \in \mathbb{Z}^k$, one has $\operatorname{diam}\{x, T_1^n x, T_2^n x, \dots, T_k^n x\} = \operatorname{diam}\{T^u y, T^u T_1^n y, T^u T_2^n y, \dots, T^u T_k^n y\} < 1$. This implies that, for $v = u_1 b_1 + u_2 b_2 + \cdots + u_k b_k$, $y(v) = y(v + nb_1) = \cdots = y(v + nb_k)$, which means that the set $v + nB$ is monochromatic.

In a similar fashion one can derive Theorem 1.2 from the following dynamical result. (Cf. [B4], p. 766).

Theorem 1.8. *Let F be a finite field and let F_∞ be the direct sum of countably many copies of F. Assume that $(T_g)_{g\in F_\infty}$ is an action of F_∞ by homeomorphisms on a compact metric space X. Then for all $\varepsilon > 0$, there exist $x \in X$ and $g \in F_\infty$, $g \neq (0, 0, \dots)$ such that* $\operatorname{diam}\{T_{cg}x, c \in F\} < \varepsilon$.

We move now to dynamical formulations of Theorems 1.4 and 1.5. We start with the discussion of Szemerédi's theorem (corresponding to $d = 1$ in Theorem 1.4). Let $E \subset \mathbb{N}$ with $\bar{d}(E) > 0$. Observe that E contains a progression of the form $\{a, a+n, \dots, a+kn\}$ if and only if $E\cap(E-n)\cap\cdots\cap(E-kn) \neq \emptyset$. It is not too hard to see that Szemerédi's theorem is actually equivalent to an ostensibly stronger statement: for all $k \in \mathbb{N}$ there exists $n \in \mathbb{N}$ such that $\bar{d}(E \cap (E - n) \cap \cdots \cap (E - kn)) > 0$. This version of Szemerédi's theorem has already a detectible dynamical content (arithmetic progressions are "produced" as the result of shifting E by $n, 2n, \dots, kn$ and getting the intersection of positive upper density). This dynamical essence of Szemerédi's theorem is embodied in Furstenberg's multiple recurrence theorem (which implies Szemerédi's result):

Theorem 1.9 ([F3]). *Let T be a measure preserving transformation of a probability measure space $(X, \mathcal{B}, \mu)$. For all $k \in \mathbb{N}$ and all $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists $n \in \mathbb{N}$ such that*

$$\mu(A \cap T^{-n}A \cap T^{-2n}A \cap \cdots \cap T^{-kn}A) > 0.$$

To derive Szemerédi's theorem from Theorem 1.9 one uses *the Furstenberg's correspondence principle* which allows one to connect the dynamics in the "pseudo-dynamical" system $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}), \bar{d}, \tau)$ (where τ is the shift map: $\tau(n) = n + 1, n \in \mathbb{Z}$) with a genuine measure preserving symbolic system which can be naturally constructed given a set $E \subset \mathbb{Z}$ with $\bar{d}(E) > 0$.

Let E be a set of integers with $\bar{d}(E) > 0$ and let X be the orbital closure of $1_E \in \{0, 1\}^{\mathbb{Z}}$ under the transformation $T\colon \omega(l) \to \omega(l + 1)$, $\omega \in \{0, 1\}^{\mathbb{Z}}$. Let $C = \{\omega \in X : \omega(0) = 1\}$. One can show (see, for example, [F4], Lemma 3.17) that there exists a T-invariant Borel measure μ on X which satisfies $\mu(C) \geq \bar{d}(E)$. By Theorem 1.9 there exists $n \in \mathbb{N}$ such that $\mu(C \cap T^{-n}C \cap T^{-2n}C \cap \cdots \cap T^{-kn}C) > 0$. If $\omega \in C \cap T^{-n}C \cap T^{-2n}C \cap \cdots \cap T^{-kn}C$ then $\{\omega, T^n\omega, T^{2n}\omega, \dots, T^{kn}\omega\} \in C$, which implies $\omega(0) = \omega(n) = \omega(2n) = \cdots = \omega(kn)$. Since ω belongs to the orbital closure of 1_E, there is an $m \in \mathbb{Z}$ such that the sequences $\omega(l)$ and $T^m 1_E(l)$ coincide for

$0 \le l \le kn$. This implies $1_E(m) = 1_E(m+n) = 1_E(m+2n) = \cdots = 1_E(m+kn)$ and gives a progression $\{m, m+n, \dots, m+kn\} \subset E$.

Similar considerations (involving Furstenberg's correspondence principle for $\mathbb{Z}^d$- and F_∞-actions) allow one to derive Theorems 1.4 and 1.5 from the following dynamical theorems.

Theorem 1.10 ([FK1]). *Let $(X, \mathcal{B}, \mu)$ be a probability measure space. For any finite set $\{T_1, T_2, \dots, T_k\}$ of commuting measure-preserving transformations of X and for all $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists $n \in \mathbb{N}$ such that*

$$\mu(A \cap T_1^{-n} A \cap T_2^{-n} A \cap \cdots \cap T_k^{-n} A) > 0.$$

Theorem 1.11. *Let $(T_g)_{g \in F_\infty}$ be a measure preserving action of $V_F = F_\infty$ on a probability measure space $(X, \mathcal{B}, \mu)$. Let $A \in \mathcal{B}$, $\mu(A) > 0$. Then for some $g \in F_\infty$, $g \neq (0, 0, \dots)$, one has $\mu\big(\bigcap_{c \in F} T_{cg} A\big) > 0$.*[4]

Remark. It is not hard to derive from Theorem 1.11 (by simple iterating procedure) the following fact: for all $k \in \mathbb{N}$ there exist $g_1, \dots, g_k \in F_\infty$ such that $\dim(\mathrm{span}\{g_1, g_2, \dots, g_k\}) = k$ and $\mu\big(\bigcap_{i=1}^k \bigcap_{c \in F} T_{cg_i} A\big) > 0$. It is this result, which, via the appropriate version of Furstenberg's correspondence principle, implies Theorem 1.5.

We will briefly discuss now the dynamical content of Hindman's theorem. Let $p \in \beta\mathbb{N}$ be an idempotent ultrafilter. The relation $p = p + p$ implies (via (1)) that a set $C \subset \mathbb{N}$ is p-large (i.e. belongs to p) if and only if $\{n \in \mathbb{N} : (C - n) \in p\} \in p$. In other words, if $p \in \beta\mathbb{N}$ is an idempotent, then every p-large set has the property that, for p-many $n \in \mathbb{N}$, the shifted set $C - n$ is also p-large. This, in turn, means that, if C is p-large then for p-many n, the set $C \cap (C - n)$ is p-large. This can be interpreted as a version of the Poincaré recurrence theorem for idempotent ultrafilter measures. The fact that the "shifting" n can itself be chosen to belong to the p-large set C comes as a bonus which, as we will presently see, immediately leads to a short and streamlined proof of Hindman's theorem.[5] Fix an idempotent ultrafilter p and let $\mathbb{N} = \bigcup_{i=1}^r C_i$ be an arbitrary finite partition of $\mathbb{N}$. Since p is a finitely additive probability measure, one (and only one) of C_i, call it C, will be p-large. As we have seen above, we can choose $n_1 \in C$ so that the set $C_1 = C \cap (C - n_1)$ is in p. We can now choose $n_2 \in C_1$ so that $n_2 > n_1$ and $C_2 = C_1 \cap (C_1 - n_2) = C \cap (C - n_1) \cap (C - n_2) \cap (C - (n_1 + n_2))$ is in p. Continuing in this fashion we will obtain an infinite set $E = \{n_1, n_2, \dots\}$ such that $FS(E) \subset C$, which concludes the proof of Hindman's theorem.

The sets of finite subsets which appear in Hindman's theorem are called in ergodic Ramsey theory *IP sets* (for infinite-dimensional parallelepiped, a term coined by H. Furstenberg and B. Weiss[6]). The notion of IP set makes sense in any commutative

[4] Theorem 1.11 follows, for example, from Theorems 1.13 and 2.9 below.

[5] The original proof of Hindman's theorem in [Hi] was very complicated. See [HiS], pp. 102–103 for references to other proofs and interesting historical comments.

[6] As we have seen, IP also naturally connects to IdemPotent.

semigroup and it is often convenient to think of IP sets as generalized semigroups. Many recurrence and convergence theorems which deal with semigroup actions can be extended to actions "along" IP sets, which leads to significant strengthening (and unification) of known results. As an illustration, we will formulate now IP versions of Theorems 1.7 and 1.10.

An $\mathcal{F}$*-sequence* in an arbitrary space Y is a sequence $(y_\alpha)_{\alpha\in\mathcal{F}}$ indexed by the set $\mathcal{F}$ of the finite nonempty subsets of $\mathbb{N}$. If Y is a commutative and multiplicative semigroup, one says that an $\mathcal{F}$-sequence defines an *IP-system* if for any $\alpha = \{i_1, i_2, \dots, i_m\} \in \mathcal{F}$ one has $y_\alpha = y_{i_1} y_{i_2} \dots y_{i_m}$. Note that if $\alpha \cap \beta = \emptyset$, then $y_{\alpha\cup\beta} = y_\alpha y_\beta$. This "partial" semigroup property turns out to be sufficient to guarantee the validity of the following multiple recurrence result which simultaneously extends Theorems 1.7 and 1.8. (It was proved first in [FW]. For a shorter proof based on the idea from [BPT], see [B2] and [B4], Cor. 2.3.)

Theorem 1.12. *If X is a compact metric space and G a commutative group of its homeomorphisms, then for any k IP-systems $(T_\alpha^{(1)})_{\alpha\in\mathcal{F}}, \dots, (T_\alpha^{(k)})_{\alpha\in\mathcal{F}}$ in G, and all $\varepsilon > 0$, there exists $\alpha \in \mathcal{F}$ and $x \in X$ such that* $\operatorname{diam}\{x, T_\alpha^{(1)}x, T_\alpha^{(2)}x, \dots, T_\alpha^{(k)}x\} < \varepsilon$.

It is clear that Theorem 1.12 implies Theorem 1.7 (any $\mathbb{Z}$ action can be viewed as a special case of an IP-system). But it is also not hard to see that Theorem 1.12 implies Theorem 1.8. See [B4], pp. 765–766 for details.

IP sets can be conveniently utilized to measure the abundance of configurations which are studied in Ramsey theory. Call a set E in a commutative semigroup G an IP* set if E has nonempty intersection with every IP set in G. It is not hard to show that every IP* set is *syndetic*.[7] To see this, one has to observe that the complement of a non-syndetic set has to contain arbitrarily long intervals, and it is not hard to show that any such set contains an IP set.

The advantage of IP* sets over syndetic sets is that the family of IP* sets has the finite intersection property (this can be shown with the help of Hindman's theorem). It follows that Theorem 1.12 not only gives a simultaneous extension of Theorems 1.7 and 1.8, but also refines each of them. For example, it follows from Theorem 1.12 that for any finite partition $\mathbb{N} = \bigcup_{i=1}^r C_i$, one of C_i has the property that, for any $k \in \mathbb{N}$, the set

$$R_k = \{d \in \mathbb{N} : \text{ for some } m \in \mathbb{N}, \{m, m+d, m+2d, \dots, m+kd\} \subset C_i\}$$

is IP*. We will see in the next section that this set R_k has much stronger intersectivity properties. (For example, R_k has nontrivial intersection with the set of values of any integer-valued polynomial $p(n)$ satisfying $p(0) = 0$).

The following powerful ergodic IP Szemerédi theorem obtained by H. Furstenberg and Y. Katznelson in [FK2] is a natural measure preserving analogue of Theorem 1.12.

[7] A subset S in a discrete semigroup G is called syndetic if finitely many translates of S cover G. If G is not commutative one has to distinguish between the notions of left and right syndetic. A left translate of S is defined as $x^{-1}S = \{g \in G : xg \in S\}$ and a right translate is defined as $Sx^{-1} = \{g \in G : gx \in S\}$.

Theorem 1.13 (See [FK2], Theorem A). *Let $(X, \mathcal{B}, \mu)$ be a probability space and G an abelian group of measure-preserving transformations of X. For all $k \in \mathbb{N}$, any IP-systems $(T_\alpha^{(1)})_{\alpha\in\mathcal{F}}, (T_\alpha^{(2)})_{\alpha\in\mathcal{F}}, \ldots, (T_\alpha^{(k)})_{\alpha\in\mathcal{F}}$ in G and all $A \in \mathcal{B}$ with $\mu(A) > 0$ there exists $\alpha \in \mathcal{F}$ such that*

$$\mu(A \cap T_\alpha^{(1)}A \cap T_\alpha^{(2)}A \cap \cdots \cap T_\alpha^{(k)}A) > 0.$$

Since the notion of an IP-system of commuting invertible measure preserving transformations generalizes the notion of a measure preserving action of a countable abelian group, Theorems 1.10 and 1.11 are immediate corollaries of Theorem 1.13. It also follows that, on a combinatorial level, Theorem 1.13 implies Theorems 1.4 and 1.5. However, Theorem 1.13 gives more! For example, it follows from it that the sets of configurations always to be found in "large" sets in $\mathbb{Z}^d$ or F_∞ are abundant in the sense that their parameters form IP* sets. These IP* versions of combinatorial results can be derived, with the help of an appropriate version of Furstenberg's correspondence principle, from the following corollary of Theorem 1.13.

Theorem 1.14. *Let $(X, \mathcal{B}, \mu)$ be a probability space, and let G be a countable Abelian group. For all $k \in \mathbb{N}$ and any measure preserving actions $(T_g^{(1)})_{g\in G}$, $(T_g^{(2)})_{g\in G}, \ldots, (T_g^{(k)})_{g\in G}$ of G on $(X, \mathcal{B}, \mu)$ and any $A \in \mathcal{B}$ with $\mu(A) > 0$, the set*

$$\{g \in G : \mu(A \cap T_g^{(1)}A \cap T_g^{(2)}A \cap \cdots \cap T_g^{(k)}A) > 0\}$$

is an IP set in G (and in particular, is syndetic).*

We will conclude this section by discussing two more classical results of Ramsey theory – the Hales–Jewett partition theorem and its density version proved in [FK3].

Consider the following generalization of tic-tac-toe: r players are taking turns in placing the symbols $s_1, \ldots, s_r$ in the $k \times k \times \cdots \times k$ (n times) array which one views as the n^{th} cartesian power A^n of a k-element set $A = \{a_1, a_2, \ldots, a_k\}$. (In the classical tic-tac-toe, we have $r = 2, k = 3, n = 2$). We are going to define now the notion of a *combinatorial line* in A^n. Identify the elements of A^n with the set $W_n(A)$ of words of length n over the alphabet A. Let $\tilde{A} = A \cup \{t\}$ be an extension of the alphabet A obtained by adding a new symbol t, and let $W_n(t)$ be the set of words of length n over $\tilde{A}$ in which the symbol t occurs. Given a word $w(t) \in W_n(t)$ let us define a combinatorial line as a set $\{w(a_1), w(a)2, \ldots, w(a_k)\}$ obtained by substituting for t the elements of A. For example, the word $43t241t2$ over the alphabet $\{1, 2, 3, 4, 5\} \cup \{t\}$ gives rise to the combinatorial line

$$\{43124112,\ 43224122,\ 43324132,\ 43424142,\ 43524152\}.$$

It is convenient to think of symbols $s_1, \ldots, s_r$ as colors; the goal of the players (in our slightly modified tic-tac-toe) is to obtain a monochromatic combinatorial line. The following theorem of Hales and Jewett [HaJ] implies that for fixed r, k and large enough n, the first player can always win.

Theorem 1.15. *Let $k, r \in \mathbb{N}$. There exists $c = c(k, r)$ such that if $n \geq c$, then for any r-coloring of the set $W_n(A)$ of words of length n over the k-letter alphabet $A = \{a_1, a_2, \ldots, a_k\}$, there is a monochromatic combinatorial line.*

One of the signs of the fundamental nature of the Hales–Jewett theorem is that one can easily derive from it its multidimensional version. For $m \in \mathbb{N}$, let $t_1, t_2, \ldots, t_m$ be m distinct variables, and let $w(t_1, t_2, \ldots, t_m)$ be a word of length n over the alphabet $A \cup \{t_1, t_2, \ldots, t_m\}$. (We assume that the letters t_i do not belong to A). If $w(t_1, t_2, \ldots, t_m)$ is a word of length n over $A \cup \{t_1, t_2, \ldots, t_m\}$ in which all of the variables $t_1, t_2, \ldots, t_m$ occur, the result of the substitution

$$\{w(t_1, t_2, \ldots, t_m)\}_{(t_1,t_2,\ldots,t_m)\in A^m} = \{w(a_{i_1}, a_{i_2}, \ldots, a_{i_m}) : a_{i_j} \in A,\ j = 1, 2, \ldots, m\}$$

is called a combinatorial m-space. Observe now that if we replace the original alphabet A by A^m, then a combinatorial line in $W_n(A^m)$ can be interpreted as a combinatorial m-space in $W_{nm}(A)$. Thus we have the following ostensibly stronger theorem as a corollary of Theorem 1.15.

Theorem 1.16. *Let $r, k, m \in \mathbb{N}$. There exists $c = c(r, k, m)$ such that if $n \geq c$, then for any r-coloring of the set $W_n(A)$ of words of length n over the k-letter alphabet A, there exists a monochromatic combinatorial m-space.*

The Hales–Jewett theorem is truly one of the cornerstones of Ramsey theory. Not only many familiar partition results such as Theorems 1.1 and 1.2 are immediate corollaries of the Hales–Jewett theorem,[8] but this result is the natural basis of many further generalizations, some of which we will encounter in the next sections. Also, the Hales–Jewett theorem and its various generalizations are utilized in proofs of various multiple recurrence results in measure preserving dynamics. For example, the Hales–Jewett theorem is used in [FK2] in the proof of Theorem 1.13. While the Hales–Jewett theorem was originally proved in a purely combinatorial way, it can also be proved with the help of Stone–Čech compactifications and by using the tools of topological dynamics. Each of these additional proofs leads in its turn to further useful results and ramifications. See [BL2], [BL3], [BBHi], [C].

In anticipation of the discussion in the next section, we are going to formulate two more versions of the Hales–Jewett theorem. Let $\mathcal{F}_0$ denote the set of finite (potentially empty) subsets of $\mathbb{N}$.

Given $k \in \mathbb{N}$, write $\mathcal{F}_0^k$ for the set of k-tuples of sets form $\mathcal{F}_0$. Let us call any $(k + 1)$ element subset of $\mathcal{F}_0^k$ of the form

$$\{(\alpha_1, \alpha_2, \ldots, \alpha_k), (\alpha_1 \cup \gamma, \alpha_2, \ldots, \alpha_k),$$
$$(\alpha_1, \alpha_2 \cup \gamma, \ldots, \alpha_k), \ldots, (\alpha_1, \alpha_2, \ldots, \alpha_k \cup \gamma)\}$$

a *simplex*.

[8] To see, for example, that Theorem 1.1 follows from the Hales–Jewett theorem, take A to be a finite field F. Then $W_n(F) = F^n$ has the natural structure of an n-dimensional vector space over F. It is easy to see that, in this case, a combinatorial m-space is an affine m-dimensional subspace of F^n.

Theorem 1.17. *For any $k \in \mathbb{N}$ and any finite coloring (= partition) of $\mathcal{F}_0^k$, there exists a monochromatic simplex.*

Here is a dynamical counterpart of Theorem 1.17.

Theorem 1.18. *Let (X, ρ) be a compact metric space. For $k \in \mathbb{N}$, let $T_{(\alpha_1,\alpha_2,\ldots,\alpha_k)}$, $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \mathcal{F}_0^k$ be a family of continuous self-maps of X such that for any $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, $(\beta_1, \beta_2, \ldots, \beta_k) \in \mathcal{F}_0^k$ satisfying $\alpha_i \cap \beta_i = \emptyset$, $i = 1, 2, \ldots, k$, one has*

$$T_{(\alpha_1\cup\beta_1,\alpha_2\cup\beta_2,\ldots,\alpha_k\cup\beta_k)} = T_{(\alpha_1,\alpha_2,\ldots,\alpha_k)} T_{(\beta_1,\beta_2,\ldots,\beta_k)}.$$

Then for all $\varepsilon > 0$ and for all $x \in X$ there exist a nonempty finite set γ and a k-tuple $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \mathcal{F}_0^k$ such that $\alpha_i \cap \gamma = \emptyset$, $i = 1, 2, \ldots, k$ and

$$\operatorname{diam}\{T_{(\alpha_1,\alpha_2,\ldots,\alpha_k)}x, T_{(\alpha_1\cup\gamma,\alpha_2,\ldots,\alpha_k)}x, T_{(\alpha_1,\alpha_2\cup\gamma,\ldots,\alpha_k)}x, \ldots, T_{(\alpha_1,\alpha_2,\ldots,\alpha_k\cup\gamma)}x\} < \varepsilon.$$

See [B1] for more discussion of various equivalent forms of the Hales–Jewett theorem and ([BL2], Proposition L) for a proof via topological dynamics.

2. Ramsey theory and multiple recurrence along polynomials

We start this section with the formulation of the Furstenberg–Sárközy theorem which has interesting links with spectral theory, Diophantine approximations, combinatorics, and dynamical systems.

Theorem 2.1 ([F4], [Sa]). *Let $E \subset \mathbb{N}$ be a set of positive upper density, and let $p(n) \in \mathbb{Z}[n]$ be a polynomial with $p(0) = 0$. Then there exist $x, y \in E$ and $n \in \mathbb{N}$ such that $x - y = p(n)$.*

This result is quite surprising. While it is not hard to show that the set of differences $E - E = \{x - y : x, y \in E\}$ of a set E with $\bar{d}(E) > 0$ is syndetic, there is, a priori, no obvious reason for the set $E - E$ to be so "well spread" as to nontrivially intersect the set of values of every polynomial $p(n) \in \mathbb{Z}[n]$ which vanishes at zero.[9] The following dynamical counterpart of Theorem 2.1, from which Theorem 2.1 follows with the help of Furstenberg's correspondence principle, is just as striking.

Theorem 2.2. *For any invertible measure preserving system $(X, \mathcal{B}, \mu, T)$, any $A \in \mathcal{B}$ with $\mu(A) > 0$, and any polynomial $p(n) \in \mathbb{Z}[n]$ with $p(0) = 0$, there exists $n \in \mathbb{N}$ such that $\mu(A \cap T^{p(n)}A) > 0$.*

The following result obtained in [BL1] gives a simultaneous generalization of Theorem 2.1 and of the Furstenberg–Katznelson multidimensional Szemerédi theorem, Theorem 1.4.

[9] T. Kamae and M. Mendès-France have shown in [KM] that a polynomial $p(n) \in \mathbb{Z}[n]$ has this "intersectivity" property if and only if $\{p(n) : n \in \mathbb{Z}\} \cap a\mathbb{Z} \neq \emptyset$ for all $a \in \mathbb{N}$. For example, the polynomial $p(n) = (n^2 - 13)(n^2 - 17)(n^2 - 221)$ is such.

Theorem 2.3 (cf. [BL1], Theorem B′). *Let $r, l \in \mathbb{N}$ and let $P\colon \mathbb{Z}^r \to \mathbb{Z}^l$ be a polynomial mapping satisfying $P(0) = 0$. For all $S \subset \mathbb{Z}^l$ with $\bar{d}(S) > 0$ and for all finite sets $F \subset \mathbb{Z}^r$, there exists elements $n \in \mathbb{N}$ and $n \in \mathbb{Z}^l$ such that $u + P(nF) = \{u + P(nx_1, nx_2, \ldots, nx_r) : (x_1, x_2, \ldots, x_r) \in F\} \subset S$.*

The ergodic theoretic result from which Theorem 2.3 is derived in [BL1] involves products of commuting measure preserving transformations evaluated at "polynomial times":

Theorem 2.4 ([BL1], Theorem A). *Let, for some $t, k \in \mathbb{N}$, $p_{1,1}(n), \ldots, p_{1,t}(n)$, $p_{2,1}(n), \ldots, p_{2,t}(n), \ldots, p_{k,1}(n), \ldots, p_{k,t}(n)$ be polynomials with rational coefficients taking integer values on the integers and satisfying $p_{i,j}(0) = 0$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, t$. Then for all probability space $(X, \mathcal{B}, \mu)$, all commuting invertible measure preserving transformations $T_1, T_2, \ldots, T_t$ of X and all $A \in \mathcal{B}$ with $\mu(A) > 0$, one has*

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{n-1} \mu(A \cap \prod_{j=1}^{t} T_j^{-p_{1,j}(n)} A \cap \prod_{j=1}^{t} T_j^{-p_{2,j}(n)} A \cap \cdots \cap \prod_{j=1}^{t} T_j^{-p_{k,j}(n)} A) > 0.$$

As we have seen in the previous section, the "linear" multiple recurrence results admit far reaching IP refinements. This leads to the question whether similar IP extensions may be obtained for polynomial results as well. For example, one would like to know whether, for every invertible measure preserving system $(X, \mathcal{B}, \mu, T)$, every polynomial $p(n) \in \mathbb{Z}[n]$ with $p(0) = 0$, and every $A \in \mathcal{B}$ with $\mu(A) > 0$, the set $\{n \in \mathbb{Z} : \mu(A \cap T^{p(n)} A) > 0\}$ is an IP* set. In other words, one would like to know whether for every infinite set $\{n_1, n_2, \ldots\} \subset \mathbb{N}$ there exists $\alpha \in \mathcal{F}$ such that $\mu(A \cap T^{p(n_\alpha)} A) > 0$, where, as in Section 1, the IP set $(n_\alpha)_{\alpha\in\mathcal{F}}$ is defined by $n_\alpha = \sum_{i\in\alpha}, \alpha \in \mathcal{F}$. The answer turns out to be yes and can be obtained by considering, instead of the conventional ergodic averages $\frac{1}{N}\sum_{n=1}^{N} \mu(A \cap T^{p(n)} A)$, the limits along IP sets (or, alternatively, limits along idempotent ultrafilters). However, a much more important novelty which is encountered when one deals with IP analogues of polynomial recurrence theorems is that one has now a bigger family of functions, namely the IP polynomials, for which the IP versions of familiar theorems make sense.

Let $q(t_1, \ldots, t_k) \in \mathbb{Z}[t_1, \ldots, t_k]$ and let $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$, $i = 1, 2 \ldots, k$, be IP sets in $\mathbb{Z}$. Then $q(\alpha) = q(n_\alpha^{(1)}, n_\alpha^{(2)}, \ldots, n_\alpha^{(k)})$ is an example of an IP polynomial. For example, if $\deg q(t_1 \ldots, t_k) = 2$, the $q(\alpha)$ will typically look like

$$q(\alpha) = \sum_{i=1}^{s} n_\alpha^{(i)} m_\alpha^{(i)} + \sum_{i=1}^{r} k_\alpha^{(i)}.$$

The following result obtained in [BFM] gives an IP extension of Theorem 2.2 for the case of several commuting transformations.

Theorem 2.5 (cf. [BFM], Corollary 2.1). *Suppose that $(X, \mathcal{B}, \mu)$ is a probability space and that $\{T_1, T_2, \dots, T_t\}$ is a collection of commuting invertible measure preserving transformations of X. Suppose that $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}} \subset \mathbb{N}$ are IP sets, $i = 1, 2, \dots, k$, and that $p_j(x_1, \dots, x_k) \in \mathbb{Z}[x_1, \dots, x_k]$ satisfy $p_j(0, 0, \dots, 0) = 0$ for $j = 1, 2, \dots, t$. Then for all $A \in \mathcal{B}$, and all $\varepsilon > 0$, there exist $\alpha \in \mathcal{F}$ such that*

$$\mu\Big(A \cap \prod_{i=1}^{t} T_i^{p_i(n_\alpha^{(1)}, n_\alpha^{(2)}, \dots, n_\alpha^{(k)})} A\Big) \geq (\mu(A))^2 - \varepsilon.$$

The next natural step is to try to extend Theorem 2.3 to a multiple recurrence result. The following IP polynomial Szemerédi theorem, obtained in [BM2] is an IP extension of Theorem 2.3.

Theorem 2.6 ([BM2], Theorem 0.9). *Suppose we are given t commuting invertible measure preserving transformations $T_1, \dots, T_t$ of a probability space $(X, \mathcal{B}, \mu)$. Let $k, r \in \mathbb{N}$ and suppose that $p_{i,j}(n_1, \dots, n_k) \in \mathbb{Q}[n_1, \dots, n_k]$ satisfy $p_{i,j}(\mathbb{Z}^k) \subset \mathbb{Z}$ and $p_{i,j}(0, 0, \dots, 0) = 0$ for $1 \leq i \leq t$, $1 \leq j \leq r$. Then for every $A \in \mathcal{B}$ with $\mu(A) > 0$, the set*

$$\big\{(n_1, \dots, n_k) \in \mathbb{Z}^k : \mu\big(\textstyle\bigcap_{j=1}^{r} \big(\prod_{i=1}^{t} T_i^{p_{i,j}(n_1,\dots,n_k)} A\big)\big) > 0\big\}$$

is an IP set in $\mathbb{Z}^k$.*

The following corollary of Theorem 2.6 can be viewed as an IP refinement of Theorem 2.5.

Theorem 2.7. *Assume that $P\colon \mathbb{Z}^r \to \mathbb{Z}^l$, $r, l \in \mathbb{N}$ is a polynomial mapping satisfying $P(0) = 0$ and let $F \subset \mathbb{Z}^r$ be a finite set. Then for all $E \subset \mathbb{Z}^l$ with $\bar{d}(E) > 0$ and all IP sets $(n_\alpha^{(i)})_{\alpha\in\mathcal{F}}$, $i = 1, \dots, r$, there exist $u \in \mathbb{Z}^l$ and $\alpha \in \mathcal{F}$ such that*

$$\{u + P(n_\alpha^{(1)} x_1, n_\alpha^{(2)} x_2, \dots, n_\alpha^{(r)} x_r) : (x_1, \dots, x_r) \in F\} \subset S.$$

We would like to mention the two combinatorial facts which play a decisive role in the proof of Theorem 2.6. The first is the Milliken–Taylor theorem ([M], [T]) which was also utilized in the proof of Theorem 2.5. The second is the polynomial Hales–Jewett theorem obtained via topological dynamics in [BL2]. The following formulation of the polynomial Hales–Jewett theorem should be juxtaposed with the formulation of its "linear" case given in Theorem 1.17.

Theorem 2.8. *For $k, d \in \mathbb{N}$, let $\mathcal{F}_0^k(\mathbb{N}^d)$ denote the set of k-tuples of finite (possibly empty) subsets of $\mathbb{N}^d = \mathbb{N} \times \cdots \times \mathbb{N}$ (d times). For every finite coloring of $\mathcal{F}_0^k(\mathbb{N}^d)$ there exists a monochromatic simplex of the form*

$$\{(\alpha_1, \alpha_2, \dots, \alpha_k), (\alpha_1 \cup \gamma^d, \alpha_2, \dots, \alpha_k),$$
$$(\alpha_1, \alpha_2 \cup \gamma^d, \dots, \alpha_k), \dots, (\alpha_1, \alpha_2, \dots, \alpha_k \cup \gamma^d)\},$$

where γ is a finite nonempty subset of $\mathbb{N}$ and $\alpha_i \cap \gamma^d = \emptyset$ for all $i = 1, 2, \dots, k$.

The polynomial Hales–Jewett theorem plays also a crucial role in the proof of the following recent result of a polynomial nature obtained in [BLM].

Theorem 2.9. *Let V, W be finite dimensional vector spaces over a countable field, let T be a measure preserving action of W on a probability measure space $(X, \mathcal{B}, \mu)$ and let $p_1, \ldots, p_k$ be polynomial mappings $V \to W$ with zero constant term. Then for all $A \in \mathcal{B}$ with $\mu(A) > 0$ there exists $c > 0$ such that the set*

$$\{v \in V : \mu(\textstyle\bigcap_{i=1}^{k} T_{(p_i(v))} A) > c\}$$

is syndetic in V.

Corollary 2.10. *Let $p_1, \ldots, p_k$ be polynomials with integer coefficients and zero constant term. For all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that whenever F is a field with $|F| \geq N$ and $E \subset F$ with $\frac{|E|}{|F|} \geq \varepsilon$, there exist $v \in F$, $v \neq 0$, and $w \in E$ such that $p_i(v) \neq 0$, $i = 1, \ldots, k$, and $\{w, w + p_1(v), \ldots, w + p_k(v)\} \subset E$ for all $i = 1, \ldots, k$.*

A dynamical counterpart of Theorem 2.8 can be formulated in direct analogy to Theorem 1.8 which corresponds to the "linear" case. (See [B1], Section 4 and [BL2] for more discussion on various equivalent forms of the polynomial Hales–Jewett theorem). Rather than dealing here with the full-fledged dynamical version, we are going to formulate a rather general corollary of the polynomial Hales–Jewett theorem which is suggestive of a further, nilpotent, generalization to be discussed in the next section.

A mapping $\mathcal{P}$ from $\mathcal{F}_0$ into a commutative (semi)group G is an *IP polynomial* of degree 0 if $\mathcal{P}$ is constant, and, inductively, is an *IP polynomial of degree $\leq d$* if for all $\beta \in \mathcal{F}_0$ there exists an IP polynomial $\mathcal{D}_\beta \mathcal{P} \colon \mathcal{F}_0(\mathbb{N} \setminus \beta) \to G$ of degree $\leq d-1$ (where $\mathcal{F}_0(\mathbb{N} \setminus \beta)$ is the set of finite subsets of $\mathbb{N} \setminus \beta$) such that $\mathcal{P}(\alpha \cup \beta) = \mathcal{P}(\alpha) + (\mathcal{D}_\beta \mathcal{P})(\alpha)$ for every $\alpha \in \mathcal{F}_0$ with $\alpha \cap \beta = \emptyset$. (One can easily check that the IP-systems introduced in Section 1 correspond to IP polynomials of degree 1 satisfying $\mathcal{P}(\emptyset) = 1_G$.)

Theorem 2.11 ([BL2]). *Let G be an abelian group of self-homeomorphisms of a compact metric space (X, ρ), let $k \in \mathbb{N}$ and let $\mathcal{P}_1, \ldots, \mathcal{P}_k$ be IP-polynomials mapping $\mathcal{F}_0$ into G and satisfying $\mathcal{P}_1(\emptyset) = \cdots = \mathcal{P}_k(\emptyset) = 1_G$. For all $\varepsilon > 0$ there exist $x \in X$ and a nonempty $\alpha \in \mathcal{F}_0$ such that $\rho(\mathcal{P}_i(\alpha)x, x) < \varepsilon$ for $i = 1, \ldots, k$.*

We conclude this section by formulating a conjecture about a density version of the polynomial Hales–Jewett theorem which would extend the partition results from [BL2], the Furstenberg–Katznelson density version of the "linear" Hales–Jewett theorem, as well as Theorems 2.4 and 2.9. For $q, d, N \in \mathbb{N}$, let $M_{q,d,N}$ be the set of q-tuples of subsets of $\{1, 2, \ldots, N\}^d$:

$$M_{q,d,N} = \{(\alpha_1, \ldots, \alpha_q) : \alpha_i \subset \{1, 2, \ldots, N\}^d, i = 1, 2, \ldots, q\}.$$

Conjecture 2.12. For all $q, d \in \mathbb{N}$ and $\varepsilon > 0$ there exists $c = c(q, d, \varepsilon)$ such that if $N > c$ and a set $S \subset M_{q,d,N}$ satisfies $\frac{|S|}{|M_{q,d,N}|} > \varepsilon$, then S contains a "simplex" of the form

$$\{(\alpha_1, \alpha_2, \dots, \alpha_q), (\alpha_1 \cup \gamma^d, \alpha_2, \dots, \alpha_q), \\ (\alpha_1, \alpha_2 \cup \gamma^d, \dots, \alpha_q), \dots, (\alpha_1, \alpha_2, \dots, \alpha_q \cup \gamma^d),$$

where $\gamma \subset \mathbb{N}$ is a nonempty set and $\alpha_i \cap \gamma^d = \emptyset$ for all $i = 1, \dots, q$.

3. Ergodic Ramsey theory in a noncommutative setting

The Ramsey theoretical results surveyed in the previous sections deal with commutative (semi)groups. One may wonder whether these results extend to noncommutative structures. A similar question suggests itself with respect to dynamics: is it true that multiple recurrence results such as, say, Theorems 1.10 and 2.4 hold true if the involved transformations do not necessarily commute? It turns out that many of the partition and density theorems (as well as their dynamical counterparts) that we encountered above do hold for nilpotent groups. On the other hand, the analogous results fail quite dramatically for solvable groups of exponential growth. (See Theorem 3.7 below).

We now formulate a nilpotent version of the polynomial Hales–Jewett theorem (see Theorems 2.8 and 2.11), from which one can derive nilpotent extensions of various abelian theorems. In order to do so we have to extend first the notion of a polynomial mapping $\mathcal{P}\colon \mathcal{F}_0 \to G$ (discussed at the end of the previous section) to a nilpotent setup.

If G is an abelian group, one can show (see [BL2], Theorem 8.3) that a mapping $\mathcal{P}\colon \mathcal{F}_0 \to G$ is an IP polynomial of degree $\leq d$ with $P(\emptyset) = 1_G$ if and only if there exists a family $\{g_{j_1,\dots,j_d}\}_{(j_1,\dots,j_d)\in\mathbb{N}^d}$ of elements of G such that for all $\alpha \in \mathcal{F}_0$ one has $\mathcal{P}(\alpha) = \prod_{(j_1,\dots,j_d)\in\alpha^d} g_{j_1,\dots,j_d}$. This characterization of IP polynomials makes sense in the nilpotent setup as well. Given a nilpotent group G, let us call a mapping $\mathcal{P}\colon \mathcal{F}_0 \to G$ an IP polynomial if for some $d \in \mathbb{N}$ there exists a family $\{g_{j_1,\dots,j_d}\}_{(j_1,\dots,j_d)\in\mathbb{N}^d}$ of elements of G and a linear order $<$ on $\mathbb{N}^d$ such that, for any $\alpha \in \mathcal{F}_0$, one has $\mathcal{F}\,P(\alpha) = \prod^{<}_{(j_1,\dots,j_d)\in\alpha^d} g_{j_1,\dots,j_d}$ (the entries in the product $\prod^{<}$ are multiplied in accordance with the order $<$). The following nilpotent version of the polynomial Hales–Jewett theorem which was obtained in [BL3] contains many of the above partition results as special cases.

Theorem 3.1 ([BL3], Theorem 0.24). *Let G be a nilpotent group of homeomorphisms of a compact metric space (X, ρ) and let $\mathcal{P}_1, \dots, \mathcal{P}_k\colon \mathcal{F}_0 \to G$ be IP polynomials satisfying $\mathcal{P}_1(\emptyset) = \cdots = \mathcal{P}_k(\emptyset) = 1_G$. Then, for all $\varepsilon > 0$, there exist $x \in X$ and a nonempty $\alpha \in \mathcal{F}_0$ such that $\rho(\mathcal{P}_i(\alpha)x, x) < \varepsilon$ for all $i = 1, 2, \dots, k$.*

One of the corollaries of Theorem 3.1 is the following nilpotent version of Theorem 1.12.

Theorem 3.2 ([BL3], Theorem 0.13). *Let G be a nilpotent group of homeomorphisms of a compact metric space (X, ρ), let $k \in \mathbb{N}$ and let $g_j^{(i)} \in G$, $i = 1, \dots, k$, $j \in \mathbb{N}$. For all $\varepsilon > 0$, there exist an element x in X and a nonempty finite subset α of $\mathbb{N}$ such that $\rho\big(\prod_{j\in\alpha} g_j^{(i)} x, x\big) < \varepsilon$ for all $i = 1, \dots, k$.*

Theorem 3.1 implies also the following results, both of which can be viewed as nilpotent extensions of Theorem 1.2.

Theorem 3.3 (cf. [BL3], Theorem 0.16). *Let $q \in \mathbb{N}$ and let G be the multiplicative group of $q \times q$ upper triangular matrices with unit diagonal over an infinite field of finite characteristic. For any finite coloring of G and any $c \in \mathbb{N}$ there exists a subgroup H of G of nilpotency class q and of cardinality $\geq c$, such that for some $h \in G$ the coset hH is monochromatic.*

Theorem 3.4 ([BL3], Theorem 0.17). *Let $q \in \mathbb{N}$ and p be a prime, with $p > q$. Let G be an infinite free q-step nilpotent group with torsion p. For any finite coloring of G and any $c \in \mathbb{N}$ there exists a free q-step nilpotent subgroup $H \subset G$ of cardinality $|H| \geq c$, such that, for some $h \in G$, the coset hH is monochromatic.*

The following theorem obtained by A. Leibman in [L1] is a nilpotent extension of Theorem 2.4, from which one can also derive a nilpotent generalization of Theorem 2.3.

Theorem 3.5 (cf. [L1], Theorem NM). *Let $k, t, r \in \mathbb{N}$. Assume G is a nilpotent group of measure preserving transformations of a probability measure space $(X, \mathcal{B}, \mu)$. Let $p_{ij}(n_1, \dots, n_k) \in \mathbb{Z}[n_1, \dots, n_k]$ with $p_{ij}(\mathbb{Z}^k) \subset \mathbb{Z}$ and $p_{ij}(0, 0, \dots, 0) = 0$, $1 \leq i \leq t$, $1 \leq j \leq r$. Then for every $A \in \mathcal{B}$ with $\mu(A) > 0$ and any $T_1, T_2, \dots, T_t \in G$, the set*

$$\Big\{(n_1, \dots, n_k) \in \mathbb{Z}^k : \mu\Big(\bigcap_{j=1}^{r} \Big(\prod_{i=1}^{t} T_i^{p_{ij}(n_1,\dots,n_k)} A\Big)\Big) > 0\Big\}$$

is syndetic.

There is every reason to believe that nilpotent versions of Theorems 2.6 and 2.9 also hold. The following conjecture, if true, will contain these and many other nilpotent results as special cases.

Conjecture 3.6. Let G be a nilpotent group of measure preserving transformations of a probability measure space $(X, \mathcal{B}, \mu)$, and let $\mathcal{P}_1, \dots, \mathcal{P}_k \colon \mathcal{F}_0 \to G$ be IP polynomials. Then for all $A \in \mathcal{B}$ with $\mu(A) > 0$ there exists a nonempty $\alpha \subset \mathbb{N}$ such that $\mu(A \cap \mathcal{P}_1(\alpha)A \cap \cdots \cap \mathcal{P}_k(\alpha)A) > 0$.

Theorem 3.5 raises question whether the assumptions can be further relaxed and whether, in particular, an analogue of Theorem 3.5 holds true if the measure preserving transformations $T_1, T_2, \dots, T_k$ generate a solvable group. Note that every finitely

generated solvable group is either of exponential growth or is virtually nilpotent, i.e. it contains a nilpotent group of finite index. (See, for example, [Ro]). Since Theorem 3.5 easily extends to virtually nilpotent groups, the question boils down to solvable groups of exponential growth. The following result answers this question in the negative, in a strong way.

Theorem 3.7 ([BL4], Theorem 1.1 (A)). *Assume that G is a finitely generated solvable group of exponential growth. There exists a measure preserving action $(T_g)_{g\in G}$ of G on a probability measure space $(X, \mathcal{B}, \mu)$, elements $g, h \in G$, and a set $A \in \mathcal{B}$ with $\mu(A) > 0$) such that $T_{g^n} A \cap T_{h^n} A = \emptyset$ for all $n \neq 0$.*

It is of interest to know to which extent the property of growth of the acting group alone is responsible for the validity of the positive and negative results formulated above. It was R. Grigorchuk who constructed in [Gri] a large family of groups of *intermediate growth*, which occupy an intermediate place between the groups of polynomial and exponential growth.

Question 3.8. Which of the above results extend to Grigorchuck's groups?

4. Generalized polynomials and dynamical systems on nilmanifolds

As we have seen in the previous section, the nilpotent framework is a natural (and often, ultimate) setup for multiple recurrence and combinatorial applications thereof. It also turns out that dynamical systems on nilmanifolds[10] are indispensable in solving problems which, on the face of it, have purely abelian character. For example, it is shown in the work of Host and Kra ([HK1]) and Ziegler ([Z]) that one can reduce the problem of establishing the existence of the L^2 limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n} x) \dots f_k(T^{kn} x),$$

where T is an invertible measure preserving transformation of a probability space $(X, \mathcal{B}, \mu)$ and $f_i \in L^\infty(X)$, to the study of the special case where (X, T) is a nilsystem. It also turns out that polynomial sequences of nilrotations (see [L2], [L3], [L5]) form an adequate setup for extending Host–Kra's and Ziegler's results to polynomial situations, that is to establishing the existence of the L^2-limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^{p_1(n)} x) \dots f_k(T^{p_k(n)} x),$$

[10] A nilmanifold is a compact homogeneous space X of a nilpotent Lie group G; a *nilrotation* is a translation of X by an element $g \in G$, $x \mapsto gx$; a *nilsystem* is a pair (X, T) where X is a nilmanifold and T is a nilrotation on X.

where the p_i are integer-valued polynomials. See [HK2] and [L4].

Another example, pertaining to recurrence, is given by the following result from [BHKR], the proof of which crucially uses the facts about nilsystems.

Theorem 4.1 ([BHKR]). *For every invertible ergodic probability measure preserving system $(X, \mathcal{B}, \mu, T)$, all $A \in \mathcal{B}$ and all $\varepsilon > 0$, the sets*

$$\{n : \mu(A \cap T^n A \cap T^{2n} A) \geq \mu(A)^3 - \varepsilon\}$$

and

$$\{n : \mu(A \cap T^n A \cap T^{2n} A \cap T^{3n} A) \geq \mu(A)^4 - \varepsilon\}$$

are syndetic.

On the other hand, there exists an ergodic system $(X, \mathcal{B}, \mu, T)$ such that for every integer $l > 1$ there exists a set $A = A(l) \in \mathcal{B}$ with $\mu(A) > 0$ and $\mu(A \cap T^n A \cap T^{2n} A \cap T^{3n} A \cap T^{4n} A) \leq \frac{1}{2}\mu(A)^l$.

We will describe now one more "nilpotent connection" recently established in [BL5]. The main object of study in [BL5] is the class of *generalized polynomials*, that is, functions obtained from conventional polynomials of one or several variables by applying the operations of addition, multiplication, and that of taking the integer part. Various classes of generalized polynomials naturally appear in diverse mathematical contexts, ranging from symbolic dynamics and mathematical games to Weyl's theorem on equidistribution[11] and recent work of Green and Tao [GreT] on arithmetic progressions in primes.[12]

Before formulating a general result from [BL5] which links generalized polynomials with nilsystems, let us briefly review a dynamical approach, due to Furstenberg, to the proof of Weyl's equidistribution theorem (see [F4], [F2]). Let $p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k = b_0 + b_1 x + b_2 \binom{x}{2} + \cdots + b_k \binom{x}{k} \in \mathbb{R}[x]$. Consider the following affine transformation, called a *skew product*, of the k-dimensional torus $\mathbb{T}^k = \mathbb{R}^k/\mathbb{Z}^k$:

$$\tau(y_1, t_2, \ldots, y_k) = (y_1 + b_k, y_2 + y_1 + b_{k-1}, \ldots, y_k + y_{k-1} + b_1).$$

Let $y = (0, \ldots, 0, b_0) \in \mathbb{T}^k$. One can check by induction that $(\tau^n y)_k = \{p(n)\}$. If a_k is irrational, the system $(\mathbb{T}^k, \tau)$ is *uniquely ergodic* (with the unique τ-invariant measure being the Lebesgue measure on $\mathbb{T}^k$) which implies (the one-dimensional version of) Weyl's theorem. (For details, see [F4], Chapter 3, Section 3.)

One can also view the skew product transformation τ as a nilrotaion. Indeed, let G be the group of upper triangular matrices $\begin{pmatrix} 1 & \alpha_{1,2} & \alpha_{1,3} & \ldots & \alpha_{1,k} \\ 0 & 1 & \alpha_{2,3} & \ldots & \alpha_{2,k} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & a_{k,k+1} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$, where $a_{i,j} \in \mathbb{Z}$ for

[11] Weyl's theorem (or rather the most quotable special case of it) says that if p is a real polynomial with at least one coefficient other than the constant term irrational then the sequence $\{p(n)\} = p(n) - [p(n)], n \in \mathbb{N}$ is uniformly distributed in the unit interval

[12] See, for example, [Gre], p. 13

$1 \le i < j \le k, a_{i,k+1} \in \mathbb{R}$ for $1 \le i < k\}$, and let Γ be the subgroup of G consisting of the matrices with integer entries. Then G is a nilpotent (non-connected) Lie group with $X = G/\Gamma \cong \mathbb{T}^k$, and the system defined on X by the nilrotation by the element $g = \begin{pmatrix} 1 & 0 & 0 & \dots & b_k \\ 0 & 1 & 0 & \dots & b_{k-1} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & b_1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \in G$ is isomorphic to the dynamical system on $\mathbb{T}^k$ defined by the skew product τ.

The following result obtained in [BL5] says, roughly, that not only generalized polynomials of the special form $\{p(n)\} = p(n) - [p(n)]$, but *any* bounded generalized polynomial can be "read off" of a nilmanifold.

Theorem 4.2. *For all $d \in \mathbb{N}$ and all bounded generalized polynomials $p\colon \mathbb{Z}^d \to \mathbb{R}$ there exists a compact nilmanifold X, an ergodic $\mathbb{Z}^d$ action $(T_n)_{n\in\mathbb{Z}^d}$ by nilrotations on X, a Riemann integrable function f on X and a point $x \in X$ such that for all $n \in \mathbb{Z}^d$ one has $p(n) = f(T_n x)$.*

Here is one of the numerous corollaries of Theorem 4.2:

Theorem 4.3. *Let $k \in \mathbb{N}$, let $U_1, \dots, U_k$ be commuting unitary operators on a Hilbert space and let $p_1, \dots, p_k$ be generalized polynomials $\mathbb{Z}^d \to \mathbb{Z}$. For any Følner sequence*[13] *$(\Phi_N)_{N=1}^{\infty}$ in $\mathbb{Z}^d$ the sequence*

$$\frac{1}{|\Phi_N|} \sum_{n\in\Phi_N} U_1^{p_1(n)} \dots U_k^{p_k(n)}$$

is convergent in the strong operator topology.

Theorem 4.3 leads to the following conjecture.

Conjecture 4.4. Theorem 4.3 remains true if the operators $U_1, \dots, U_k$ appearing in its formulation generate a nilpotent group.

Assume now that the unitary operators $U_1, \dots, U_k$ are induced by commuting measure preserving transformations $T_1, \dots, T_k$ acting on a probability space $(X, \mathcal{B}, \mu)$. In this case it is natural to inquire under which conditions on the generalized integer-valued polynomials p_i one has

$$\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T_1^{p_1(n)} T_2^{p_2(n)} \dots T_k^{p_k(n)} A) > 0$$

for all $A \in \mathcal{B}$ with $\mu(A) > 0$. In the case of conventional integer-valued polynomials a satisfactory sufficient condition for positivity of the above limit is that $p_i(0) = 0$,

[13] A sequence $(\Phi_N)_{N=1}^{\infty}$ of finite subsets of a (countable) group G is called (left) *Følner* if for all $g \in G$, $|g\Phi_N \cap \Phi_N|/|\Phi_N| \to 1$ as $N \to \infty$. In $\mathbb{Z}^d$ a common choice of Følner sequence is a sequence of parallelepipeds $\Phi_N = \prod_{i=1}^d [a_{N,i}, b_{N,i}]$ with $b_{N,i} - a_{N,i} \to \infty$ as $N \to \infty$ for all $i = 1, \dots, d$.

$i = 1, \ldots, k$; the following conjecture extends this fact to generalized polynomials. Let us denote by $\mathcal{P}_0$ the set of generalized polynomials which can be constructed (with the help of addition, multiplication, and taking of integer part) from conventional polynomials with *zero constant term*.

Conjecture 4.5. Let $k \in \mathbb{N}$ and $p_1, \ldots, p_k \in \mathcal{P}_0$. Then for any commuting invertible measure preserving transformations $T_1, \ldots, T_k$ of a probability measure space $(X, \mathcal{B}, \mu)$ and all $A \in \mathcal{B}$ with $\mu(A) > 0$ one has

$$\lim_{N-M\to\infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T_1^{p_1(n)} T_2^{p_2(n)} \ldots T_k^{p_k(n)} A) > 0.$$

Note that Conjecture 4.5 implies that $\{n : \mu(A \cap T_1^{p_1(n)} T_2^{p_2(n)} \ldots T_k^{p_k(n)} A) > 0\}$ is a syndetic set. This, in turn, is a special case of the following conjecture, which extends the polynomial Szemerédi theorem (cf. Theorem 2.4 above) to generalized polynomials belonging to $\mathcal{P}_0$.

Conjecture 4.6. Let $(X, \mathcal{B}, \mu)$ be a probability measure space, let $k, r \in \mathbb{N}$, let $T_1, \ldots, T_k$ be commuting invertible measure preserving transformations of X and let $p_{i,j} \in \mathcal{P}_0$, $i = 1, \ldots, k$, $j = 1, \ldots, r$. Then for all $A \in \mathcal{B}$ with $\mu(A) > 0$, the set

$$\{n \in \mathbb{Z} : \mu(A \cap T_1^{p_{1,1}(n)} \ldots T_k^{p_{k,1}(n)} A \cap \cdots \cap T_1^{p_{1,r}(n)} \ldots T_k^{p_{k,r}(n)} A) > 0\}$$

is syndetic in $\mathbb{Z}$.

5. Amenable groups and ergodic Ramsey theory

It was John von Neumann who, in his study of the Hausdorff–Banach–Tarski paradox, introduced a class of group which nowadays are called *amenable* and which are widely recognized as providing the natural context for ergodic theory. In particular, many classical notions and results pertaining to 1-parameter group actions extend naturally to amenable groups. (See for example [OW] and [Li]). As we will see in this section, countable amenable groups also form a natural framework for Furstenberg's correspondence principle and hence for ergodic Ramsey theory.

Definition 5.1. A semigroup G is amenable if there exists an *invariant mean* on the space $B(G)$ of real-valued bounded functions on G, that is, a positive linear function $L\colon B(G) \to \mathbb{R}$ satisfying

(i) $L(\mathbf{1}_G) = 1$.

(ii) $L(f_g) = L({}_g f) = L(f)$ for all $f \in B(G)$ and $g \in G$, where $f_g(t) := f(tg)$ and ${}_g f(t) := f(gt)$.

The existence of an invariant mean is only one item from a long list of equivalent properties , such as the characterization of amenability given in the next theorem, some of which are far from being obvious and, moreover, are valid for groups (or special classes of semigroups) only. (See, for example, [W], Theorem 10.11.) The following theorem was established by Følner in [Fø]. (See also [N] for a simplified proof.)

Theorem 5.2. *A countable group G is amenable if and only if it has a left Følner sequence, namely a sequence of finite sets $\Phi_n \subset G$, $n \in \mathbb{N}$, with $|\Phi_n| \to \infty$ and such that for all $g \in G$, $\frac{|g\Phi_n \cap \Phi_n|}{|\Phi_n|} \to 1$ as $n \to \infty$.*

While there seems to be no general method of constructing a Følner sequence in an amenable group defined, say, by generators and relations, in many concrete, especially abelian situations, one has no problem finding a Følner sequence. For example, the parallelepipeds mentioned in footnote 13 and the sets $F_n \subset F_\infty$ defined in section 1 form natural Følner sequences in $\mathbb{Z}^d$ and in F_∞, respectively.

Before moving to discuss the Ramsey-theoretical aspects of amenable groups we want to mention that while the class of amenable groups is quite rich (in particular it contains all solvable and locally finite groups), it does not contain such classical groups as $\mathrm{SL}(n, \mathbb{Z})$ for $n \geq 2$.

Given a countable amenable group G and a left Følner sequence $(\Phi_n)_{n\in\mathbb{N}}$, one can define the upper density of a set $E \subset G$ with respect to $(\Phi_n)_{n\in\mathbb{N}}$ by $\bar{d}_{(\Phi_n)}(E) = \limsup_{n\to\infty} \frac{|E\cap\Phi_n|}{|\Phi_n|}$. Note that it immediately follows from the definition of a left Følner sequence that for all $g \in G$ and $E \subset G$ one has $\bar{d}_{(\Phi_n)}(gE) = \bar{d}_{(\Phi_n)}(E)$. By analogy with some known results about sets of positive density in abelian or nilpotent groups, one can expect that large sets in G, i.e. sets having positive upper density with respect to some Følner sequence, will contain some nontrivial configurations. The known results support this point of view and lead to a natural conjecture which will be formulated at the end of this section.

We formulate now a version of Furstenberg's correspondence principle for countable amenable groups.

Theorem 5.3 (See [B2] , Theorem 6.4.17). *Let G be a countable amenable group and assume that $E \subset G$ has positive upper density with respect to some left Følner sequence $(\Phi_n)_{n\in\mathbb{N}}$: $\bar{d}_{(\Phi_n)}(E) > 0$. Then there exists a probability measure preserving system $(X, \mathcal{B}, \mu, (T_g)_{g\in G})$ and a set $A \in \mathcal{B}$ with $\mu(A) = \bar{d}_{(\Phi_n)}(E)$ such that for all $k \in \mathbb{N}$ and $g_1, \ldots, g_k \in G$ one has*

$$\bar{d}_{(\Phi_n)}(E \cap g_1 E \cap \cdots \cap g_k E) \geq \mu(A \cap T_{g_1} A \cap \cdots \cap T_{g_k} A).$$

Remark. One can extend Theorem 5.3 to general countable amenable semigroups if instead of using Følner sequences (which cannot always be found in amenable semigroups) one defines a set $E \subset G$ to be large if for some left-invariant mean L on $B(G)$ one has $L(1_E) > 0$. (See [BM1], Theorem 2.1.)

As an illustration of the usefulness of amenable considerations, let us consider the (abelian and cancellative) semigroup $(\mathbb{N}, \cdot)$. Let

$$S_n = \{p_1^{i_1} p_2^{i_2} \cdots p_n^{i_n}, 0 \le i_j \le n, 1 \le j \le n\},$$

where p_i, $i = 1, 2, \ldots$, are primes in arbitrary order. It is not hard to show that for any sequence of positive integers $(a_n)_{n\in\mathbb{N}}$, the sets $a_n S_n$, $n \in \mathbb{N}$ form a Følner sequence in $(\mathbb{N}, \cdot)$.

Definition 5.4. A set $E \subset \mathbb{N}$ is called *multiplicatively large* if for some Følner sequence $(\Phi_n)_{n\in\mathbb{N}}$ in $(\mathbb{N}, \cdot)$ one has $\overline{d}_{(\Phi_n)}(E) > 0$.

Notice that the notions of additive and multiplicative largeness which are defined via Følner sets in, respectively, $(\mathbb{N}, +)$ and $(\mathbb{N}, \cdot)$ are different. For example the set O of odd natural numbers has additive density $\frac{1}{2}$ with respect to every Følner sequence in $(\mathbb{N}, +)$, while O has zero density with respect to every Følner sequence in $(\mathbb{N}, \cdot)$. In the other direction, consider for example a Følner sequence $(a_n S_n)_{n\in\mathbb{N}}$ in $(\mathbb{N}, \cdot)$, where the S_n are defined above and the a_n satisfy $a_n > |S_n|$. Then the set $E = \bigcup_{n=1}^{\infty} a_n S_n$ has zero additive density with respect to every Følner sequence in $(\mathbb{N}, +)$, while E has multiplicative density 1 with respect to the Følner sequence $(a_n S_n)_{n\in\mathbb{N}}$.

As may be expected by mere analogy with additively large sets, multiplicatively large sets always contain (many) geometric progressions. (This can be derived, for example, with the help of the IP Szemerédi theorem, Theorem 1.13 above). It turns out, however, that multiplicatively large sets also contain some other, somewhat unexpected *geoarithmetic* configurations.

Theorem 5.5 (See [B5], Theorems 3.11 and 3.15). *Let $E \subset \mathbb{N}$ be a multiplicatively large set. For all $k \in \mathbb{N}$, there exist $a, b, c, d, e, q \in \mathbb{N}$ such that $\{q^i(a + id) : 0 \le i, j \le k\} \subset E$ and $\{b(c + ie)^j : 0 \le i, j \le k\} \subset E$.*

We conclude this section (and this survey) by addressing the question about possible amenable extensions of the multiple recurrence results. While it is not clear how even to formulate an amenable generalization of the one-dimensional Szemerédi theorem, it is not too hard to guess what should be an amenable version of the multidimensional Szemerédi theorem!

Let G be a group and $k \in \mathbb{N}$. Let us call a $(k + 1)$-element set in the cartesian product G^k a *simplex* if it is of the form

$$\{(a_1, a_2, \ldots, a_k), (ga_1, a_2, \ldots, a_k), (ga_1, ga_2, \ldots, a_k), \ldots, (ga_1, ga_2, \ldots, ga_k)\}$$

for some $a_1, \ldots, a_k, g \in G$, and denote it by $S(a_1, \ldots, a_k; g)$. The following conjecture is known for $k = 2$. (See [BMZ], Theorem 6.1.)

Conjecture 5.6. Let $k \in \mathbb{N}$ and suppose that G is a countable amenable group. Assume that a set $E \subset G^k$ has positive upper density with respect to some Følner

sequence in G^k. Then the set

$$\{g \in G : \text{there exist } (a_1, \dots, a_k) \in G^k \text{ such that } S(a_1, \dots, a_k; g) \subset E\}$$

is syndetic in G.

References

[B1] Bergelson, V., Ergodic Ramsey theory - an update. In *Ergodic Theory of $\mathbb{Z}^d$-actions* (Warwick, 1993-1994), London Math. Soc. Lecture Note Ser. 228, Cambridge University Press, Cambridge 1996, 273–296.

[B2] Bergelson, V., Ergodic theory and diophantine problems. In *Topics in Symbolic Dynamics and Applications* (Temuco 1997), London Math. Soc. Lecture Note Ser. 277, Cambridge University Press, Cambridge 2000, 167–205.

[B3] Bergelson, V., Minimal idempotents and ergodic Ramsey theory. In *Topics in Dynamics and Ergodic Theory*, London Math. Soc. Lecture Note Ser. 279, Cambridge University Press, Cambridge 2000, 167–205.

[B4] Bergelson, V., Combinatorial and diophantine applications of ergodic theory. Appendix A by A. Leibman and Appendix B by A. Quas and M. Wierdl. In *Handbook of Dynamical Systems* (ed. by B. Hasselblatt and A. Katok), Vol. 1B, Elsevier, Amsterdam 2006, 745–841.

[B5] Bergelson, V., Multiplicatively large sets and ergodic Ramsey theory. *Israel J. Math.* **148** (2005), 23–40.

[BBHi] Bergelson, V., Blass, A., Hindman, N., Partition theorems for spaces of variable words. *Proc. London Math. Soc.* (3) **68** (3) (1994), 449–476.

[BFM] Bergelson, V., Furstenberg, H., McCutcheon, R., IP-sets and polynomial recurrence. *Ergodic Theory Dynam. Systems* **16** (5) (1996), 963–974.

[BHKR] Bergelson, V., Host, B., Kra, B., Multiple recurrence and nilsequences. (With an appendix by I. Ruzsa.) *Invent. Math.* **160** (2) (2005), 261–303.

[BL1] Bergelson, V., Leibman, A., Polynomial extensions of Van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.* **9** (1996), 725–753.

[BL2] Bergelson, V., Leibman, A., Set-polynomials and polynomial extension of the Hales-Jewett theorem. *Ann. of Math.* (2) **150** (1) (1999), 33–75.

[BL3] Bergelson, V., Leibman A., Topological multiple recurrence for polynomial configurations in nilpotent groups. *Adv. Math.* **175** (2) (2003), 271–296.

[BL4] Bergelson, V., Leibman, A., Failure of the Roth theorem for solvable groups of exponential growth. *Ergodic Theory Dynam. Systems*. **24** (2004), 45–53.

[BL5] Bergelson, V., Leibman, A., Distribution of values of bounded generalized polynomials. Submitted.

[BLM] Bergelson, V., Leibman A., McCutcheon, R., Polynomial Szemerédi theorems for countable modules over integral domains and finite fields. *J. Anal. Math.* **95** (2005) 243–296.

[BM1] Bergelson, V., McCutcheon, R., Recurrence for semigroup actions and a non-commutative Schur Theorem. In *Topological Dynamics and Applications* (Minneapolis, MN, 1995), Contemp. Math. 215, Amer. Math. Soc., Providence, RI, 1998, 205–222.

[BM2] Bergelson, V., McCutcheon, R., An ergodic IP polynomial Szemerédi theorem. *Mem. Amer. Math. Soc.* **146** (695) (2000).

[BMZ] Bergelson, V., McCutcheon, R., Zhang, Q., A Roth theorem for amenable groups. *Amer. J. Math.* **119** (6) 1997, 1173–1211.

[BPT] Blaszczyk, A., Plewik, S., Turek, S., Topological multidimensional van der Waerden theorem. *Comment. Math. Univ. Carolin.* **30** (4) (1989), 783–787.

[C] Carlson, T., Some unifying principles in Ramsey theory. *Discrete Math.* **68** (1988), 117–169.

[CN] Comfort, W., Negrepontis, S., *The Theory of Ultrafilters*. Grundlehren Math. Wiss. 211, Springer-Verlag, Berlin 1974.

[E] Ellis, R., A semigroup associated with a transformation group. *Trans. Amer. Math. Soc.* **94** (1960), 272–281.

[Fø] Følner, E., On groups with full Banach mean values. *Mat. Scand.* **3** (1955), 243–354.

[F1] Furstenberg, H., *Stationary processes and prediction theory*. Ann. of Math. Stud. 44, Princeton University Press, Princeton, N.J., 1960

[F2] Furstenberg, H., The structure of distal flows. *Amer. J. Math.* **85** (1963), 477–515.

[F3] Furstenberg, H., Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Anal. Math.* **31** (1977), 204–256.

[F4] Furstenberg, H., *Recurrence in Ergodic Theory and Combinatorial Number Theory*. Princeton University Press, Princeton, N.J., 1981.

[FK1] Furstenberg, H., Katznelson, Y., An ergodic Szemerédi theorem for commuting transformations. *J. Anal. Math.* **34** (1978), 275–291.

[FK2] Furstenberg, H., Katznelson, Y., An ergodic Szemerédi theorem for IP-systems and combinatorial theory. *J. Anal. Math.* **45** (1985), 117–168.

[FK3] Furstenberg, H., Katznelson, Y., A density version of the Hales-Jewett theorem. *J. Anal. Math.* **57** (1991), 64–119.

[FW] Furstenberg, H., Weiss, B., Topological dynamics and combinatorial number theory. *J. Anal. Math.* **34** (1978), 61–85.

[GraRS] Graham, R., Rothschild, B., Spencer, J., *Ramsey Theory*. Wiley, New York 1980.

[Gre] Green, B. Long arithmetic progressions of primes. Available at http://arxiv.org/abs/math.NT/0508063.

[GreT] Green, B., Tao, T., The primes contain arbitrarily long arithmetic progressions. *Ann. of Math.*, to appear.

[Gri] Grigorchuk, R. I., Degrees of growth of finitely generated groups and the theory of invariant means. *Izv. Akad. Nauk SSSR, Ser 1 Mat.* **48** (1984) (5) 939–985; English transl. *Math. USSR-Izv.* **25** (2) (1985), 259–300.

[HaJ] Hales, A. W., Jewett, R. I. Regularity and positional games. *Trans. Amer. Math. Soc.* **106** (1963), 222–229.

[Hi] Hindman, N., Finite sums from sequences within cells of a partition of $\mathbb{N}$. *J. Combin. Theory* (*Ser. A*) **17** (1974), 1–11.

[HiS] Hindman, N., Strauss, D., *Algebra in the Stone-Čech compactification. Theory and Applications.* De Gruyter Exp. Math. 27, Walter de Gruyter, Berlin 1998.

[HK1] Host, B., Kra, B., Nonconventional ergodic averages and nilmanifolds. *Ann. of Math.* **161** (2005) (1), 397–488.

[HK2] Host, B., Kra, B., Convergence of polynomial ergodic averages. *Israel J. of Math.* **148** (2005), 267–276.

[KM] Kamae, T., Mendès-France, M., Van der Corput's difference theorem. *Israel J. Math.* **31** (3-4) (1978), 335–342.

[L1] Leibman, A., Multiple recurrence theorem for measure preserving actions of a nilpotent group. *Geom. Funct. Anal.* **8** (1998), 853–931.

[L2] Leibman, A., Pointwise convergence of ergodic averages for polynomial sequences of transformations on a nilmanifold. *Ergodic Theory Dynam. Systems* **25** (2005), 201–213.

[L3] Leibman, A., Pointwise convergence of ergodic averages for polynomial actions of $\mathbb{Z}^d$ by translations on a nilmanifold. *Ergodic Theory Dynam. Systems* **25** (2005), 215–225.

[L4] Leibman, A. Convergence of multiple ergodic averages along polynomials of several variables. *Israel J. Math.* **146** (2005), 303–315.

[L5] Leibman, A., Orbits on nilmanifolds under the actions of polynomial sequences of translations. Preprint; http://www.math.ohio-state.edu/~leibman/preprints/.

[Li] Lindenstrauss, E., Pointwise theorems for amenable groups. *Invent. Math.* **146** (2001), 259–295.

[M] Milliken, K., Ramsey's theorem with sums or unions. *J. Combin. Theory* (*Ser. A*) **18** (1975), 276–290.

[N] Namioka, I., Følner's condition for amenable semi-groups. *Math. Scand.* **15** (1964), 18–28.

[O] Ornstein, Donald S., Bernoulli shifts with the same entropy are isomorphic. *Adv. Math.* **4**, 337-352.

[OW] Ornstein, D., Weiss, B., Entropy and isomorphism theorems for actions of amenable groups. *J. Anal. Math.* **48** (1987), 1–141.

[R] Rado, R., Note on combinatorial analysis. *Proc. London Math. Soc.* **48** (1993), 122–160.

[Ro] Rosenblatt, J., Invariant measures and growth conditions. *Trans. Amer. Math. Soc.* **193** (1974), 33–53.

[Sa] Sàrközy, On difference sets of integers III. *Acta Math. Acad. Sci. Hungar.* **31** (1978) 125–149.

[Sz] Szemerédi, E., On sets of integers containing no k elements in arithmetic progression. *Acta Arith.* **27** (1975), 199–245.

[T] Taylor, A., A canonical partition relation for finite subsets of ω. *J. Combin. Theory* (*Ser A.*) **17** (1974), 1–11.

[W] Wagon, S., *The Banach-Tarski Paradox*. Cambridge University Press, Cambridge 1985.

[Z] Ziegler, T., Universal characteristic factors and Furstenberg averages. *J. Amer. Math. Soc.*, to appear.

Department of Mathematics, Ohio State University, Columbus, OH 43210, U.S.A.

E-mail: vitaly@math.osu.edu

Hyperbolic billiards and statistical physics

Nikolai Chernov and Dmitry Dolgopyat[1]

Abstract. Mathematical theory of billiards is a fascinating subject providing a fertile source of new problems as well as conjecture testing in dynamics, geometry, mathematical physics and spectral theory. This survey is devoted to planar hyperbolic billiards with emphasis on their applications in statistical physics, where they provide many physically interesting and mathematically tractable models.

Mathematics Subject Classification (2000). Primary 37D50; Secondary 34C29, 60F.

Keywords. Hyperbolic billiards, mixing, limit theorems.

1. Introduction

Let $\mathcal{D}$ be a bounded domain on a plane or a 2D torus with piecewise smooth boundary. A billiard system in $\mathcal{D}$ is generated by a single particle moving freely inside $\mathcal{D}$ with specular reflections off the boundary $\partial\mathcal{D}$. The phase space of a billiard is a 3D manifold Ω; the corresponding flow $\Phi^t : \Omega \to \Omega$ preserves the Liouville measure μ (which is uniform on Ω). The space of all collision points makes a 2D cross-section $\mathcal{M} \subset \Omega$, and the corresponding return map $\mathcal{F} : \mathcal{M} \to \mathcal{M}$ (called billiard map) preserves a natural smooth probability measure m.

The billiard is hyperbolic if the flow Φ^t and the map $\mathcal{F}$ have non-zero Lyapunov exponents. The first class of hyperbolic billiards was introduced [86] by Sinai in 1970; he proved that if the boundary of $\mathcal{D}$ is convex inward, then the billiard is hyperbolic, ergodic, mixing and K-mixing. He called such models dispersing billiards, now they are called Sinai billiards. They are also proven to be Bernoulli [43]. A few years later Bunimovich discovered [9], [10] that billiards in some domains $\mathcal{D}$ whose boundary is convex outward are also hyperbolic, due to a special 'defocusing mechanism'; the most celebrated example of his billiards is a stadium. More general classes of planar hyperbolic billiards are described in [94], [95], [63], [41]; we refer to [48], [26] for extensive surveys on hyperbolic billiards.

Billiards differ from classical smooth hyperbolic systems (Anosov and Axiom A flows and maps) in several respects. First of all, many hyperbolic billiards have

[1]We thank our adviser Ya. G. Sinai for introducing us to this subject and constant encouragement during our work. We thank our numerous collaborators, especially P. Balint, L. A. Bunimovich, R. de la Llave, G. Eyink, J. Jebowitz, R. Markarian, D. Szasz, T. Varju, L.-S. Young, and H.-K. Zhang, for many useful discussions on the subject of this survey. NC was partially supported by NSF. DD was partially supported by NSF and IPST.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

non-uniform expansion and contraction rates (for example, if the moving particle is almost tangent to a convex outward arc of the boundary, then it will 'slide', and many reflections will occur in rapid succession during a short interval of time; a similar phenomenon occurs in a cusp on the boundary). Only dispersing billiards without cusps have uniform expansion and contraction rates.

Second, and most importantly, the billiard dynamics have singularities – phase points where both map $\mathcal{F}$ and flow Φ^t become discontinuous and have unbounded derivatives. Singularities come from two sources:

(a) Grazing collisions. In this case nearby trajectories can land on boundary components that lie far apart.

(b) Corners. In this case two nearby trajectories can hit different boundary pieces converging to a corner and get reflected at substantially different angles.

Moreover, billiards without horizon (where the length of the free path between collisions is unbounded) have infinitely many singularity curves in phase space.

Singularities in billiards lead to an unpleasant fragmentation of phase space. More precisely, any curve in unstable cones gets expanded (locally), but the singularities may cut its image into many pieces, some of them shorter than the original curve, which then will have to spend time on recovering. This makes billiards similar to non-uniformly hyperbolic systems such as quadratic maps or Henon attractors.

In [96], [97] Young has proposed two general methods for studying non-uniformly hyperbolic systems: tower method and coupling method.

The first one generalizes well-known Markov partitions ([85]). The latter divide phase space into rectangles ('building blocks') that have a direct product structure and being moved under the dynamics intersect one another in a proper (Markov) way. In the tower method only one rectangle is used and its images only need to intersect itself in the Markov way for some (not all) iterations. The tower construction is thus more flexible than that of Markov partitions, but the symbolic dynamics it provides is just as good as the one furnished by a Markov partition.

The coupling method is designed to directly control the dependence between the past and the future. Since points with the same past history form unstable manifolds, one wants to show that the images of any two curves in unstable cones have asymptotically the same distribution ([84]). To this end one partitions those curves into small subsets and pairs subsets of the first curve with those of the second one so that the images of the paired (coupled) points remain close to each other at all times (i.e. lie on the same stable manifold).

Both methods proved to be very efficient and produced many sharp results, as we describe below. We observe here that the tower method allows us to use functional analytic tools, in particular the theory of transfer operators [3], [71], which provide very precise asymptotic expansions. However the transfer operator approach requires a suitably defined space of functions (observables), which is sometimes too restrictive and dependent on the model at hand. For this reason the results obtained by the tower approach are often less explicit and the dependence on parameters of the model is less transparent. The coupling approach, being more elementary if less sophisticated,

gives more explicit bounds and makes it easier to work with several systems at a time.

Our survey is organized as follows. Section 2 describes statistical properties of dispersing billiards. Section 3 is devoted to systems with slow mixing rates. Section 4 deals with billiards in the presence of external forces and discusses transport coefficients and their dependence on parameters. Section 5 is devoted to interacting billiard particles, and Section 6 deals with infinite volume billiards.

We will denote by $\mathcal{N}(0, \sigma^2)$ a normal random variable (vector) with zero mean and variance (covariance matrix) σ^2, and by ρ_{σ^2} its density function.

2. Dispersing billiards

Dispersing billiards make the oldest and most extensively studied class of all chaotic billiards. They, arguably, have the strongest statistical properties among all billiards. We need to suppose that all corners have positively measured angles (no cusps) to guarantee uniform expansion and contraction rates.

The main difficulty in the studies of billiards is to cope with the destructive effect of fragmentation caused by singularities (we note that fragmentation may badly affect even relatively simple expanding maps so that they would fail to have good statistical properties [92]). In billiards, to cope with pathological fragmentation one imposes the following 'non-degeneracy' condition: there exist $m \in \mathbb{N}$, $\delta > 0$, and $\theta_0 < 1$ such that for any smooth unstable curve W of length less than δ

$$\sum_i \lambda_{i,m} \le \theta_0, \tag{1}$$

where the sum runs over all smooth components $W_{i,m} \subset \mathcal{F}^m(W)$ and $\lambda_{i,m}$ is the factor of contraction of $W_{i,m}$ under $\mathcal{F}^{-m}$. Roughly speaking (1) says that there no too-degenerate singularities such as multiple passages through the corners. (1) always holds if there are no corners, i.e. if $\partial\mathcal{D}$ is smooth, because for grazing collisions the expansion factor approaches infinity on one side of each singularity line, but in it is unknown if the condition (1) always holds in dispersing billiards with corners, nor if it is really necessary for the results presented below.

Let $\mathbb{B}^d_\alpha$ be the space of bounded $\mathbb{R}^d$-valued functions which are uniformly α-Hölder continuous on each component of $\mathcal{M}$ where the map $\mathcal{F}$ is smooth. We write $\mathbb{B}_\alpha$ for $\mathbb{B}^1_\alpha$. Let $\overline{\mathbb{B}}^d_\alpha = \{A \in \mathbb{B}^d_\alpha : m(A) = 0\}$. For any function $A \in \overline{\mathbb{B}}^d_\alpha$ we denote by $\sigma^2(A)$ the $d \times d$ (diffusion) matrix with components

$$\sigma^2_{ij}(A) = \sum_{n=-\infty}^{\infty} m\big(A_i\,(A_j \circ \mathcal{F}^n)\big) \tag{2}$$

(if this series converges). Denote $S_n(x) = \sum_{k=0}^{n-1} A(\mathcal{F}^k x)$.

Theorem 1. *The following four results hold under the condition* (1):

(a) (Exponential mixing [96], [18], [20]) *There is a constant* $\theta < 1$ *such that for every* $A, B \in \overline{\mathbb{B}}_\alpha$, *for all* $n \in \mathbb{Z}$

$$\left|m\big(A\,(B \circ \mathcal{F}^n)\big)\right| \le \text{const}\,\theta^{|n|},$$

which, in particular, implies the convergence of the series (2).

(b) (Functional Central Limit Theorem [11], [12], [20]) *For* $A \in \overline{\mathbb{B}}_\alpha^d$ *define a continuous function* $W_n(t)$ *by letting* $W_n(k/n) = S_k/\sqrt{n}$ *and interpolating linearly in between. Then* $W_n(t)$ *weakly converges, as* $n \to \infty$, *to a Brownian motion* (*Wiener process*) *with covariance matrix* $\sigma^2(A)$.

(c) (Almost sure invariance principle [66], [20]) *There exist* $\lambda > 0$ *such that for any* $A \in \overline{\mathbb{B}}_\alpha$ *we can find* (*after possibly enlarging the phase space*) *a Brownian motion* (*Wiener process*) $w(t)$ *with variance* $\sigma^2(A)$ *such that for almost all* x *there is* n_0 *such that for* $n \ge n_0$,

$$|S_n - w(n)| < n^{\frac{1}{2}-\lambda}.$$

(d) (Local Limit Theorem [90]) *Suppose* $A \in \overline{\mathbb{B}}_\alpha^d$ *takes values in a closed subgroup* $V \subset \mathbb{R}^d$ *of rank* r *and that there is no* $B \in L^2_m(\mathcal{M})$ *such that* $A + B - B \circ \mathcal{F}$ *belongs to a proper closed subgroup of* V. *Then for any continuous function* G *with compact support and for any sequence* $\{k_n\}$ *such that* $k_n/\sqrt{n} \to z \in \mathbb{R}^d$,

$$n^{r/2} m\big(G(S_n - k_n)\big) \to \rho_{\sigma^2(A)}(z) \int F\,dl$$

where l *is the Haar measure on* V.

Parts (a)–(c) of Theorem 1 can be proved by both tower method and coupling method ([96], [18], [20], [66]). The only known proof of part (d) uses the tower construction. It would be useful to derive the last part also by the coupling approach, since then it would be applicable to systems depending on parameters.

If $\mathcal{A}$ is a function on Ω, then standard reduction methods [73], [67] allow us to extend parts (b) and (c) to $S_t(X) = \int_0^t \mathcal{A}(\Phi^s X)\,ds$. The corresponding covariance matrix $\tilde{\sigma}^2(\mathcal{A})$ can be computed as follows. Consider the function $A(x) = \int_0^{\tau(x)} \mathcal{A}(\Phi^s x)\,ds$ on $\mathcal{M}$, where $\tau(x)$ is the length of the free path. Then

$$\tilde{\sigma}^2(\mathcal{A}) = \sigma^2(A)/\bar{\tau}, \tag{3}$$

where $\bar{\tau} = \pi\,\text{Area}(\mathcal{D})/\text{length}(\partial\mathcal{D})$ is the mean free path in the billiard system [16].

It would be also nice to extend the part (c) to multidimensional observables, as the almost sure invariance principle readily implies other limit laws – the law of iterated logarithm, integral tests, etc. [20].

Problem 1. Prove the almost sure invariance principle for $\mathbb{R}^d$ valued observables.

The above results can be applied to the Lorentz gas in $\mathbb{R}^2$. Consider a particle moving on the plane between a periodic array of fixed convex disjoint obstacles (scatterers). The natural phase space of this system is the unit tangent bundle to the plane minus the scatterers, and the natural invariant measure is infinite (σ-finite). But since the dynamics commute with the $\mathbb{Z}^2$ action we can factor the latter out and reduce the system to a dispersing billiard on the unit torus.

Let S_n be the center of the scatterer the particle hits at the nth collision. Then $S_n - S_{n-1}$ factors to a function $H(\mathcal{F}^{n-1}x)$ on the collision space $\mathcal{M}$ of the toral billiard. To apply Theorem 1 we need to assume that this billiard has finite horizon (a uniformly bounded free path), since otherwise $H(x)$ is unbounded and has infinite second moment. (This is *not* a technical restriction, the following result actually fails without the horizon assumption, see Section 3.) Let $q(t)$ be the position of the moving particle at time t.

Theorem 2. *The following five results hold for finite horizon Lorentz gases:*

(a) ([11], [12]) $S_n/\sqrt{n}$ *converges weakly to* $\mathcal{N}(0, \sigma^2)$ *where*

$$\sigma_{ij}^2 = \sum_{n=-\infty}^{\infty} m\big(H_i\,(H_j \circ \mathcal{F}^n)\big). \tag{4}$$

(b) ([11], [12]) $q(t)/\sqrt{t}$ *converges to* $\mathcal{N}(0, \sigma^2/\bar{\tau})$.

(c) ([90]) $m(S_n = 0) \sim 1/(2\pi \det(\sigma)n)$.

(d) ([30], [78]) S_n *is recurrent.*

(e) *The Lorentz gas is ergodic with respect to its* σ*-finite invariant measure.*

Parts (c) and (d) are recent. Part (e) follows from part (d) and [79].

Parts (c) and (d) indicate that S_n behaves like a random walk.

Problem 2. Extend the analogy between S_n and random walks (for instance, investigate the statistics of returns).

Some results in this direction are obtained in [40]. Results for geodesic flows on negatively curved surfaces can be found in [1].

3. Slow mixing and non-standard limit theorems

Here we describe some hyperbolic billiards with non-uniform expansion and contraction rates. Such are billiards with convex outward boundary components, semidispersing billiards (where the boundary is convex inward, but at some points its curvature vanishes, i.e. the boundary 'flattens'), as well as dispersing billiards with cusps. All these billiards have one feature in common - there are arbitrarily long series of reflections without expansion or contraction, which compromise the hyperbolicity.

Such series of 'idle' reflections occur in certain well defined regions in phase space. If $\hat{\mathcal{M}} \subset \mathcal{M}$ is their complement, then the return map $\hat{\mathcal{F}} : \hat{\mathcal{M}} \to \hat{\mathcal{M}}$ will have uniform expansion and contraction rates, so Young's methods will apply. The distribution of return times to $\hat{\mathcal{M}}$ then determines the rates of mixing:

Theorem 3. (a) ([28]) *If $\mathcal{D}$ is a Bunimovich stadium (a table with C^1 boundary consisting of two semicircles and two parallel line segments) and $A, B \in \overline{\mathbb{B}}_\alpha$, then*

$$\left|m\big(A\,(B \circ \mathcal{F}^n)\big)\right| \le \text{const} \cdot (\ln |n|)^2/|n|. \tag{5}$$

The same bound holds for modified stadia bounded by two circular arcs and two non-parallel line segments.

(b) ([28]) *If $\mathcal{D}$ is a Bunimovich billiard table bounded by several circular arcs that do not exceed semicircles an $A, B \in \overline{\mathbb{B}}_\alpha$, then*

$$\left|m\big(A\,(B \circ \mathcal{F}^n)\big)\right| \le \text{const} \cdot (\ln |n|)^3/|n|^2.$$

(c) ([29]) *Let $\mathcal{D}$ be a dispersing billiard table except the curvature of $\partial\mathcal{D}$ vanishes at two points $P, Q \in \partial\mathcal{D}$ such that the segment PQ is a periodic orbit of period two. More precisely let the boundary $\partial\mathcal{D}$ contain two curves $y = \pm(|x|^\beta + 1)$, where $\beta > 2$, so that $P = (0, 1)$ and $Q = (0, -1)$. Then for $A, B \in \overline{\mathbb{B}}_\alpha$,*

$$\left|m\big(A\,(B \circ \mathcal{F}^n)\big)\right| \le \text{const} \cdot (\ln |n|)^{a+1}/|n|^a \text{ where } a = \frac{\beta+2}{\beta-2}.$$

The logarithmic factors here are an artifact of the method used; they can presumably be removed [22] by approximating the map $\mathcal{F}$ on $\mathcal{M} \setminus \hat{\mathcal{M}}$ with a Markov chain (the region $\mathcal{M} \setminus \hat{\mathcal{M}}$ consists of countably many 'cells' that make almost a Markov partition). The bound (5) is expected for dispersing billiards with cusps [61], but this case turns out to be much harder; it is currently under investigation [27].

If correlations decay like $\mathcal{O}(1/n)$, as in Bunimovich stadia, the series (2) is likely to diverge, so the central limit theorem is likely to fail. This happens because the main contribution to the sum S_n comes from long series of (highly correlated) reflections without expansion or contraction. Again, we can employ the return map $\hat{\mathcal{F}} : \hat{\mathcal{M}} \to \hat{\mathcal{M}}$ and replace the given observable A with its 'cumulative' version

$$\bar{A}(x) = \sum_{n=0}^{R(x)-1} A(\mathcal{F}^n x), \tag{6}$$

where $\hat{\mathcal{F}}(x) = \mathcal{F}^{R(x)}(x)$, i.e. $R(x)$ is the first return time (to $\hat{\mathcal{M}}$), but such $\bar{A}$ will usually be unbounded and have heavy tails.

First studies of limit laws for observables with heavy tails were undertaken by Aaronson and Denker [2] for systems with Markov partitions. Their results were extended to non-uniformly hyperbolic maps with Young towers by Balint and Gouezel [4]; they gave an abstract criterion for convergence to a Gaussian law under a non-classical normalization (the case which is most relevant for billiards).

Balint and Gouezel [4] redefined $R(x)$ in (6) to be the first return time to the only rectangle in Young's tower and proved a limit theorem under the assumption that $\bar{A}$ has a distribution in a non-standard domain of attraction of Gaussian law. They applied this criterion to a Bunimovich stadium bounded by two semicircles of radius 1 and two line segments Γ_1 and Γ_2 of length $L > 0$ each: given a Hölder continuous observable $A \in C^\alpha(\mathcal{M})$, denote by

$$I(A) = \frac{1}{2L} \int_{\Gamma_1 \cup \Gamma_2} A(s, \boldsymbol{n})\, ds$$

its average value on the set of normal vectors $\boldsymbol{n}$ attached to Γ_1 and Γ_2. (A slower decay of correlations for the stadium, compared to other Bunimovich billiards, is caused by trajectories bouncing between two flat sides of $\mathcal{D}$ and $I(A)$ represents the contribution of such trajectories.)

Theorem 4. *The following results hold for Bunimovich stadia:*

(a) *If* $I(A) \neq 0$ *then* $S_n/\sqrt{n \ln n} \to \mathcal{N}(0, \sigma^2(A))$, *where*

$$\sigma^2(A) = \frac{4 + 3\ln 3}{4 - 3\ln 3} \times \frac{[I(A)]^2 L^2}{4(\pi + L)}. \tag{7}$$

(b) *If* $I(A) = 0$, *then there is* $\sigma_0^2 > 0$ *such that* $S_n/\sqrt{n} \to \mathcal{N}(0, \sigma_0^2)$.

As before, the approach of [67] allows us to extend this result to flows.

The abstract criterion of [4] should be applicable to a large number of systems. One of them is a periodic Lorentz gas without horizon [91]. In this case orbits which never collide with the scatterers lie in a finite number of families of corridors $\Pi_i \subset \mathbb{R}^2$. The projection of each corridor onto the torus is a strip bounded by two periodic orbits (which in general case correspond to fixed points of the collision map $\mathcal{F}$). Let w_i denote the vector joining the successive collisions along the bounding orbits for the corridor Π_i. Let also f_i denote a vector parallel to w_i but whose length equals the width of Π_i. Consider a nonnegative quadratic form

$$Q(v) = \frac{1}{\text{length}(\partial\mathcal{D})} \sum_i |w_i|\, \langle f_i, v\rangle^2.$$

This form corresponds to a 2×2 symmetric positive semidefinite matrix σ^2.

Theorem 5 ([91]). *Suppose there are at least two non-parallel corridors in a Lorentz gas without horizon. Then* $\sigma^2 > 0$ *and*

(a) $S_n/\sqrt{n \ln n} \to \mathcal{N}(0, \sigma^2)$;

(b) *If* $k_n/\sqrt{n \ln n} \to z$ *then* $n \ln n \cdot m(S_n = k_n) \to \rho_{\sigma^2}(z)$;

(c) S_n *is recurrent*;

(d) *the Lorentz gas is ergodic with respect to its* σ*-finite invariant measure.*

Problem 3. Prove a functional central limit theorem in the setting of [4].

Solving this problem would lead to a complete asymptotic description of the flight process in Lorentz gases without horizon.

4. Transport coefficients

Here we begin the discussion of billiard-related models of mathematical physics. The simplest one is a billiard $\mathcal{D}$ where the particle moves under an external force

$$\dot{v} = F(q, v). \tag{8}$$

Such systems were investigated in [19] under the assumptions that $\mathcal{D}$ is the torus with a finite number of disjoint convex scatterers and finite horizon. To prevent unlimited acceleration or deceleration of the particle, it was assumed that there was an integral of motion ("energy") $\mathcal{E}(q, v)$ such that each ray $(q, \alpha v)$, $\alpha \in \mathbb{R}_+$ intersects each level surface $\{\mathcal{E} = c\}$ in exactly one point. To preserve hyperbolicity, it was assumed that $\|F\|_{C^1}$ is small.

Such forces include potential forces ($F = -\nabla U$), magnetic forces ($F = B(q) \times v$) and electrical forces with the so-called Gaussian thermostat:

$$F = E(q) - \frac{\langle E(q), v\rangle}{\|v\|^2} v. \tag{9}$$

Fix an energy surface $\{\mathcal{E}(q, v) = \text{const}\}$ containing a point with unit speed. Under our assumptions on $\mathcal{E}$ this level surface is diffeomorphic to the unit tangent bundle Ω over $\mathcal{D}$ and the collision space $\mathcal{M}_F$ is diffeomorphic to $\mathcal{M}$. Denote by $\mathcal{F}_F \colon \mathcal{M}_F \to \mathcal{M}_F$ the corresponding return map.

Theorem 6 ([19]). *$\mathcal{F}_F$ has a unique SRB (Sinai–Ruelle–Bowen) measure m_F, i.e. for Lebesgue almost every $x \in \mathcal{M}_F$ and all $A \in C(\mathcal{M}_F)$*

$$\frac{1}{n} \sum_{i=0}^{n-1} A(\mathcal{F}_F^i x) \to \int_{\mathcal{M}_F} A \, dm_F.$$

The map $\mathcal{F}_F$ is exponentially mixing and satisfies the Central Limit Theorem (cf. Theorem 1).

As usual one can derive from this the existence (and uniqueness) of the SRB measure μ_F for the continuous time system.

Another interesting modification of billiard dynamics results from replacing the "hard core" collisions with the boundary by interaction with a"soft" potential near the boundary. We do not describe such systems here for the lack of space referring the reader to [60].

Theorem 6 implies the existence of various transport coefficients for planar Lorentz gas with finite horizon. For example, consider a thermostated electrical force (9) with a constant field $E(q) = E = \text{const}$, and let m_E denote the SRB measure on the $\{\mathcal{E} = 1/2\}$ energy surface.

Theorem 7 ([24]). *There is a bilinear form ω such that for $A \in C^\alpha(\mathcal{M})$*

$$m_E(A) = m(A) + \omega(A, E) + o(\|E\|).$$

To illustrate these results, let q_n denote the location of the particle on the plane at its nth collision, then Theorem 6 implies for almost all x the average displacement $(q_n - q_0)/n$ converges to a limit, $J(E)$, i.e. the system exhibits an electrical current. Theorem 7 implies

$$J(E) = ME + o(\|E\|) \quad \text{(Ohm's Law)},$$

where M is a 2×2 matrix, see below.

One interesting open problem is to study the dependence of the measure m_F of the force F, for example the smoothness of m_E as a function of the electrical field E. For hyperbolic maps without singularities SRB measure depends smoothly on parameters [51], [76], [77]. For systems with singularities the results and methods of [24] demonstrate that the SRB measure is differentiable at points where it has smooth densities (e.g. $E = 0$ in the previous example).

In fact there is an explicit expression for the derivative (Kawasaki formula). To state it let $\mathcal{F}_\varepsilon$ be a one-parameter family of maps such that $\mathcal{F}_0 = \mathcal{F}$ has a smooth SRB measure and for small ε the map $\mathcal{F}_\varepsilon$ has an SRB measure m_ε, too, and the convergence to the steady state m_ε, in the sense that if ν is a smooth probability measure on $\mathcal{M}$ and $A \in C^\alpha(\mathcal{M})$ then $\nu(A \circ \mathcal{F}_\varepsilon^n) \to m_\varepsilon(A)$, is exponential in n and uniform in ε. Let $X = \frac{d}{d\varepsilon}\big|_{\varepsilon=0}(\mathcal{F}_\varepsilon \circ \mathcal{F}^{-1})$. Then

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} m_\varepsilon(A) = -\sum_{n=0}^{\infty} \int_{\mathcal{M}} \operatorname{div}_m(X)\, A(\mathcal{F}^n x)\, dm(x). \tag{10}$$

For the constant electrical field E the Kawasaki formula reads $DJ|_{E=0} = \frac{1}{2}\sigma^2$, where σ^2 is defined by (4). Hence

$$J = \frac{1}{2}\sigma^2 E + o(\|E\|), \tag{11}$$

which is known in physics as Einstein relation.

On the other hand numerical experiments [8] seem to indicate that $J(E)$ is *not* smooth for $E \neq 0$. Similar lack of smoothness is observed in ([44], [45], [47]) for expanding interval maps, but the billiard case seems to be more complicated. Indeed the smoothness of SRB measures (or the lack thereof) seems to be intimately related to the dynamics of the singularity set. For 1D maps the singularity set is finite whereas for 2D maps the singularity set is one-dimensional, and so one can expect some statistics for the evolution of that set.

Problem 4. Prove that the SRB measure, as a function of parameters, is not smooth (generically). Derive relations between its Hölder exponent near a given parameter value and other dynamical invariants, such as Lyapunov exponents, entropy, etc.

A related issue is the dependence of infinite correlation sums, such as the one in (10), on the geometry of the billiard table. This issue was addressed in [23]. Given

a domain $\mathcal{D} \subset \mathbb{T}^2$, an additional round scatterer is placed in $\mathcal{D}$ with a fixed radius $R > 0$ and a (variable) center Q; then one gets a family of billiard maps $\mathcal{F}_Q$ acting on the same collision space $\mathcal{M}$ and having a common smooth invariant measure m. For any smooth functions A, B on $\mathcal{M}$ with zero mean let

$$\sigma^2_{A,B}(Q) = \sum_{n=-\infty}^{\infty} m\big(A\,(B \circ \mathcal{F}_Q^n)\big). \tag{12}$$

It is proven in [23] that $\sigma^2_{A,B}(Q)$ is a log-Lipschitz continuous function of Q:

$$\big|\sigma^2_{A,B}(Q_1) - \sigma^2_{A,B}(Q_2)\big| \le \text{const}\,\Delta \ln(1/\Delta), \quad \text{where } \Delta = \|Q_1 - Q_2\|. \tag{13}$$

Problem 5. Is (13) an optimal bound?

Problem 6. Extend the analysis of [23] to dissipative systems studied in [19].

In particular is it true that the dependence on parameters is typically more regular for conservative systems?

Problem 7. Consider the class $\mathbb{S}$ of all Sinai billiard tables on $\mathbb{T}^2$ and deform a given table $\mathcal{D}$ continuously in C^4 so that it approaches the natural boundary of $\mathbb{S}$. Investigate the limit behavior of the diffusion matrix $\sigma^2(\mathcal{D})$.

If we only consider generic boundary points of $\mathbb{S}$, then this problem splits into three subproblems:

(a) What happens when two scatterers nearly touch each other?

(b) What happens when the boundary flattens so that a periodic trajectory with nearly zero curvature appears?

(c) What happens when one of the scatterers shrinks to a point?

Analogues of Problem 7 were investigated for expanding maps [38] and for geodesic flows on negatively curved surfaces [13]. For Sinai billiards, only problem (c) has been tackled in [23], see Theorem 9 (a) below. The first step towards problem (a) is to establish mixing bounds for billiards with cusps (for problem (b) this task has been accomplished in [29], see Theorem 3 (c)).

One can also study the behavior of other dynamical invariants, such as entropy and Lyapunov exponents, see [16], [32], [48], [14].

5. Interacting particles

One may hope that after so many results have been obtained for one particle dynamics in dispersing billiards, a comparable analysis could be done for multi-particle systems, including models of statistical mechanics where the number of particles grows to infinity. However not much has been achieved up to now. Recently there has been a significant progress in the study of stochastically interacting particles [52], [93], but

the problems involving deterministic systems appear to be much more difficult. One notable result is [70] where Euler equation is derived for Hamiltonian systems with a weak noise, however that particular noise is of a very special form, and its choice remains to be justified by microscopic considerations.

Regarding models with finitely many particles, the most celebrated one is a gas of hard balls in a box with periodic boundary conditions (i.e. on a torus $\mathbb{T}^d$). The ergodicity of this system is a classical hypothesis in statistical mechanics attributed to L. Boltzmann and first mathematically studied by Sinai [83], [86], see [48]. The hyperbolicity and ergodicity for this system have been proven in fairly general cases only recently [80], [81], but a proof in full generality is not yet available.

Problem 8. Prove the ergodicity of N hard balls on a torus $\mathbb{T}^d$ for every $N \geq 3$ and $d \geq 2$ and for arbitrary masses $m_1, \ldots, m_N$ of the balls.

The existing proofs [80], [81] cover 'generic' mass vectors $\{m_1, \ldots, m_N\}$ (apart from unspecified submanifolds of codimension one in $\mathbb{R}^N$). Besides, the existing proofs heavily rely on abstract algebraic-geometric considerations, and it is important to find more explicit and dynamical arguments.

A system of N hard balls on $\mathbb{T}^d$ can be reduced to semi-dispersing billiards in a Nd-dimensional torus with a number of multidimensional cylinders removed. Now the considerations of Section 3 suggest that the rate of mixing for gases of hard balls is quite slow. Physicists estimated that correlation functions for the flow decay as $\mathcal{O}(t^{-d/2})$, see [42], [72].

Problem 9. Investigate mixing rate for gases of N hard balls in $\mathbb{T}^d$ or N hard disks on a Sinai billiard table.

An important feature of systems considered in statistical mechanics is that there are several different scales in space and time. This can complicate the study since the problem of interest tend to involve several 'levels' of parameters, but on the other hand one can expect certain simplifications; for example, Hamiltonian systems of N particles which are not ergodic (and this is, generically, the case due to the KAM theory) may behave as ergodic in the thermodynamical limit $N \to \infty$ (see e.g. [31], Chapter 9). Another example is that some pathologies slowing the mixing rates can be suppressed on large time-space scales, thus the system may behave as strongly chaotic.

A significant progress in the study of multi-scale systems with chaotic fast motion has been achieved recently, see [39] and references wherein. In this section we describe the first rigorous result on multi-scale billiard systems [23].

Consider a system of two particles moving on a 2D torus with a finite number of fixed convex scatterers (we assume that the resulting region $\mathcal{D} \subset \mathbb{T}^2$ has finite horizon). Particles collide with the scatterers and with each other elastically. The first particle called **P** is a heavy disk of mass $M \gg 1$ and radius $R \sim 1$. The second particle called **p** is a dimensionless point of unit mass.

In equilibrium, the kinetic energies of **P** and **p** are comparable, and then **P** will move practically with constant velocity, without noticing **p**. A more interesting development occurs if the initial velocity of **P** is zero. Assume that the initial speed of **p** is 1 and that its initial state is chosen randomly from the unit tangent bundle over $\mathcal{D}$. Then the position Q of **P** at time t becomes a random process $Q_M(t)$. We want to describe the motion of **P** in the interior of $\mathcal{D}$ (before it has chance to reach $\partial\mathcal{D}$), so we fix a small $\delta > 0$ and stop **P** once it comes within distance δ from $\partial\mathcal{D}$. Under a non-degeneracy condition on $\mathcal{D}$, see below, the following is proved:

Theorem 8 ([23]). *As $M \to \infty$, the process $Q_M(\tau M^{2/3})$ converges weakly to the solution of the following stochastic differential equation*

$$\ddot{Q} = \tilde{\sigma}(Q)\,\dot{w} \tag{14}$$

where $\dot{w}$ is the white noise and the 2×2 matrix $\tilde{\sigma}(Q)$ is the positive square root of

$$\tilde{\sigma}^2(Q) = \sigma^2(Q)/\bar{\tau},$$

compare this to (3); *here $\bar{\tau} = \pi(\mathrm{Area}(\mathcal{D}) - \mathrm{Area}(\mathbf{P}))/(\mathrm{length}(\partial\mathcal{D}) + \mathrm{length}(\partial\mathbf{P}))$ is the mean free path for the fast particle* **p** *and*

$$\sigma^2(Q) = \sum_{n=-\infty}^{\infty} m\big(A\,(A \circ \mathcal{F}_Q^n)^T\big)$$

where $\mathcal{F}_Q$ is defined before Eq. (12) *and $A \in \mathbb{B}^2$ is defined by* (18) *below.*

The non-degeneracy condition mentioned above is $\sigma^2(Q) > 0$ for all Q. This condition allows us to 'promote' the log-Lipschitz continuity of σ^2 given by (13) to the log-Lipschitz continuity of $\tilde{\sigma}$ and then show that the equation (14) is well posed. This illustrates the importance of Problems 5 and 6 for homogenization theory. The fact that $\sigma^2(Q)$ is non-degenerate, apart from a codimension infinity subset of $\mathbb{S}$, follows from [12].

To understand (14) observe that when **P** collides with **p** the tangential component of its velocity remains unchanged while the normal component changes as follows

$$V_{\mathrm{new}}^{\perp} = \frac{M-1}{M+1}\,V_{\mathrm{old}}^{\perp} + \frac{2}{M+1}\,v_{\mathrm{old}}^{\perp} = V_{\mathrm{old}}^{\perp} + \frac{2}{M}\,v_{\mathrm{old}}^{\perp} + \mathcal{O}\Big(\frac{1}{M^{3/2}}\Big) \tag{15}$$

where $v_{\mathrm{old}}^{\perp}$ is the normal component of the velocity of **p** (the estimate on the remainder term uses the fact that due to the energy conservation $M\|V\|^2 + \|v\|^2 = 1$ the speed of **P** never exceeds $1/\sqrt{M}$). Hence velocity of **P** after n collisions equals

$$V_n = \frac{2}{M}\sum_{i=1}^{n} v_i^{\perp} + \mathcal{O}\Big(\frac{n}{M^{3/2}}\Big) \tag{16}$$

where $v_i^\perp$ is the normal component of the velocity of **p** before the i-th collision of **P** with **p**. As we need to count all the collisions of **p**, both with **P** and $\partial\mathcal{D}$, then (16) takes form

$$V_n = \frac{2}{M}\sum_{i=1}^{n} A\circ\mathcal{F}^i + \mathcal{O}\Big(\frac{n}{M^{3/2}}\Big) \tag{17}$$

where $\mathcal{F}$ is the collision map in our system of two particles and

$$A = 2v^\perp \text{ if } \mathbf{p} \text{ collides with } \mathbf{P} \text{ and } 0 \text{ otherwise.} \tag{18}$$

As $M\to\infty$, our system approaches the limit where **P** does not move ($Q\equiv$ const) and **p** bounces off $\partial\mathcal{D}\cup\mathbf{P}$ elastically, thus its collision map is $\mathcal{F}_Q$. For this limiting system, Theorem 1 (c) says that if $n = M^\alpha d\tau$, then

$$\sum_{i=1}^{n} A\circ\mathcal{F}_Q^i \sim M^{\alpha/2}\,\sigma(Q)\,dw(\tau) \tag{19}$$

where $w(\tau)$ is the standard Brownian motion. Since $Q = \int V dt$ and the integral of the Brownian motion grows as $t^{3/2}$, it is natural to take $\alpha = 2/3$ in (19), so that $M^{3\alpha/2}/M \sim 1$, cf. (16), and expect the limiting process to satisfy (14).

In the proof of Theorem 8 we had to show that the two-particle collision map $\mathcal{F}$ in (17) could be well approximated by the limiting billiard map $\mathcal{F}_Q$ in (19). While the trajectories of individual points under these two maps tend to diverge exponentially fast, the images of curves in unstable cones tend to stay close together, and we proved this by a probabilistic version of the shadowing lemma developed in [37]. Then we decomposed the initial smooth measure into one-dimensional measures on unstable curves (each curve W with a smooth measure ν on it was called a standard pair) and adapted Young's coupling method to relate the image of each standard pair (W,ν) under the map $\mathcal{F}^n$ and that under $\mathcal{F}_Q^n$, as n grows.

The system described above is a very simplified version of the classical Brownian motion where a macroscopic particle is submerged into a liquid consisting of many small molecules. In our model the liquid is represented by a single particle, but its chaotic scattering off the walls effectively replaced the chaotic motion of the molecules coming presumably from inter-particle interactions.

One feature of Theorem 8 which may be surprising at first glance is that the diffusion matrix σ^2 is position dependent – the feature one does not expect for the classical Brownian particle. The reason is that the size of **P** is comparable to the size of the container $\mathcal{D}$, so that typical time between successive collisions of **p** with **P** is of order one, hence **p** has memory of the previous collisions with **P** giving rise to a location dependent diffusion matrix. This dependence disappears if **P** is macroscopically small (but microscopically large!):

Theorem 9 ([23]). *As $R\to 0$ we have*

$$\tilde\sigma^2(Q) = \frac{8R}{3\mathrm{Area}(\mathcal{D})}\,I + \mathcal{P}(Q)\,R^2 + o(R^2), \tag{20}$$

where $\mathcal{P}(Q)$ is a weighted Poincaré series. Furthermore, there is a function $M_0(R)$ such that if $M \to \infty$ and $R \to 0$ with $M > M_0(R)$, then $Q(\tau R^{-1/3} M^{2/3})$ converges weakly to the process

$$\sqrt{\frac{8}{3\text{Area}(\mathcal{D})}} \int_0^\tau w(s)\, ds$$

where $w(s)$ is the standard Brownian Motion.

Observe that the formula (20) would easily follow if the collisions between **p** and **P** made a random Poisson process with intensity proportional to $2R/\text{Area}(\mathcal{D})$ (the inverse of the mean intercollision time).

We remark that since we have a single fast particle **p**, its collisions with the boundary $\partial\mathcal{D}$ are the only source of chaos. If $\mathcal{D}$ is a convex smooth table, for example, then due to the presence of caustics [53] there is a positive probability that **p** and **P** will never meet, so Theorem 8 fails in that case.

Problem 10. Prove Theorems 8 and 9 for two particles in a square box.

In a square box, the fast particle may bounce off between two parallel sides for a long time without running into the disk, so the dynamics has slow mixing rates, cf. Section 3. According to the results of [4], see Theorem 4, one expects a non-standard normalization for most observables. However the observable given by (18) vanishes on $\partial\mathcal{D}$ (since the velocity of **P** does not change during the collisions of **p** with the walls), so we are actually in the context of Theorem 4 (b), hence Central Limit Theorem may hold despite the overall slow mixing rates.

The extension of Theorem 9 to a square box leads (by using a standard reflection of the box across its boundary) to a new model – a fast particle moving on a plane with a periodic configuration of identical circular scatterers of radius $R \to 0$. This system is interesting in its own rights, but not much is known about its asymptotic properties as $R \to 0$. A lot of work has been done on the case where scatters are placed at random (see [7], [82] and references wherein) but the periodic case is much more complicated, see [46]. Even the distribution of the free path is a non-trivial task accomplished only recently [6].

The results of [23] extend, without much changes, to systems with several heavy disks and one fast particle, as long as the disks do not collide with each other or with the boundary of the table (of course this restricts the analysis to a fairly short interval of time). Let us, for example, formulate an analogue of Theorem 8 in this situation. Let k be the number of heavy disks which are initially at rest. Then, after rescaling time by $\tau = M^{-2/3}t$, the velocity of the limiting process satisfies

$$\frac{d}{d\tau}\begin{pmatrix} V_1 \\ \vdots \\ V_k \end{pmatrix} = \sigma_{Q_1 \dots Q_k}\, \dot{w}(\tau),$$

where $\dot{w}$ is a standard k-dimensional white noise. Note that even though the heavy disks are not allowed to interact with each other directly, each one "feels" the presence of the others through the diffusion matrix $\sigma_{Q_1 \dots Q_k}$, which depends on the positions of all the disks.

A much more difficult problem arises if there are several fast particles.

Problem 11. Extend Theorems 8 and 9 to systems with several fast particles.

In this case the limiting ($M \to \infty$) system consists of several non-interacting particles moving on the same dispersing billiard table (the heavy disk(s) will be "frozen" as $M = \infty$). Such a system can be reduced to a semidispersing billiard in a higher dimensional container, however that billiard will have very poor statistical properties. In fact, it will not be even fully hyperbolic – several of its Lyapunov exponents corresponding to the flow directions of the particles will vanish.

A more promising strategy for this case is to deal directly with the continuous time dynamics. Then the limiting system of several non-interacting fast particles is a direct product of one-particle billiard flows. To extend the results of [23] to this model we need to generalize their methods to the continuous time setting, and we also need good estimates for mixing rates of dispersing billiard flows.

Problem 12. Estimate the decay of correlations for dispersing billiard flows.

The studies of flow correlations are notoriously difficult (the main reason is that there is no expansion or contraction in the flow direction). Even for classical Anosov flows no estimates on correlations were available until the late 1990s. Only recently various estimates were obtained on the decay of correlations for smooth uniformly hyperbolic flows [17], [35], [59]. Some of them were just extended to nonuniformly hyperbolic flows [65], including Sinai billiards: it was shown [65] that for a 'prevalent' set of Sinai billiards with finite horizon, flow correlations decay faster than any polynomial function.

We expect that the flow correlations for Sinai billiards with finite horizon actually decay exponentially fast. Moreover, it appears that a sub-optimal ('stretched exponential') bound developed in [17] can be extended to billiard flows, and this is our work in progress. With some of these estimates, albeit less than optimal, we might be able to handle the above system of several fast particles.

Interestingly, the mixing rates of the billiard flow may not match those of the billiard map. For instance, in Sinai billiards without horizon the billiard map has fast (exponential) decay of correlations [18], but the flow is apparently very slowly mixing due to long flights without collisions [5]. On the contrary, in Sinai tables with cusps, the billiard map appears to have polynomial mixing rates, see Section 3, but the flow may very well be exponentially mixing, as the particle can only spend a limited time in a cusp. The same happens in Bunimovich billiards bounded only by circular arcs that do not exceed semicircles – the billiard map has slow mixing rates (Theorem 3), but the flow is possibly fast mixing, as sliding along arcs (which slows down the collision map) does not take much flow time.

The next step toward a more realistic model of Brownian motion would be to study several light particles of a positive radius $r > 0$. (If there is only one light particle, such an extension is immediate since 'fattening' the light particle is equivalent to 'fattening' the disk **P** and the scatterers by the same width r.) It is however reasonable to assume that the light particles are much smaller than the heavy one, i.e. $r \ll R$. In this case one can presumably treat consecutive collisions as independent, so that in the limit $r \to 0$ the collision process becomes Markovian. An intermediate step in this project would be

Problem 13. Consider a system of k identical particles of radius $r \ll 1$ moving on a dispersing billiard table $\mathcal{D}$. Let $E_i(t)$ denote the energy of the ith particle at time t. Prove that the vector

$$\{E_1(\tau/r), E_2(\tau/r), \dots, E_k(\tau/r)\}$$

converges, as $r \to 0$, to a Markov process with transition probability density given by the Boltzmann collision kernel [15]. This means that if particles i and j collide so that the angles between their velocities and the normal are in the intervals $[\phi_i, \phi_i + d\phi_i]$ and $[\phi_j, \phi_j + d\phi_j]$, respectively, with intensity

$$\frac{\left|\sqrt{2E_i}\cos\phi_i - \sqrt{2E_j}\cos\phi_j\right| d\phi_i\, d\phi_j}{4\pi^2 \text{Area}(\mathcal{D})},$$

and then the particle i transfers energy $E_i\cos^2\phi_i - E_j\cos^2\phi_j$ to the particle j.

The proof should proceed as follows. As long as the particles do not interact, the evolution of the system is a direct product of dynamics of individual particles. This holds true whenever the particle centers $q_1, \dots, q_k$ are $> 2r$ units of length apart. Hence we need to investigate the statistics of visits of phase orbits to $\Delta_r = \{\min_{i\neq j} \|q_i - q_j\| \le 2r\}$, which is a set of small measure. Visits of orbits of (weakly) hyperbolic systems to small measure sets have been studied in many papers, see [36], [50] and the references wherein We observe that Theorem 9 (a) is the first step in the direction of Problem 13.

Next, recall that in Theorem 8 we did not allow the disk **P** to come too close to the boundary $\partial\mathcal{D}$; this restricted our analysis to intervals of time $t = \mathcal{O}(M^{2/3})$. During these times the speed of **P** remains small, $\|V\| = \mathcal{O}(M^{-2/3})$, thus the system is still far from equilibrium, as $M\|V\|^2 = \mathcal{O}(M^{-1/3}) \ll 1$.

Problem 14. Investigate the system of two particles **P** and **p** beyond the time of the first collision of **P** with $\partial\mathcal{D}$. In particular, how long does it take this system to approach equilibrium (where the energies of the particles become equal)?

There are two difficulties here. First, when **P** comes too close to the wall $\partial\mathcal{D}$, the mixing properties of the limiting ($M \to \infty$) billiard system deteriorate, because a narrow channel forms between **P** and the wall. Once the fast particle **p** is trapped in

that channel, it will bounce between **P** and the wall for quite a while before getting out; thus many highly correlated collisions between our particles occur, all pushing **P** in the same direction (off the wall). Thus we expect $\|\sigma(Q)\| \to \infty$ as the channel narrows. The precise rate of growth of $\|\sigma(Q)\|$ is important for the boundary conditions for equation (14), hence Problem 7 is relevant here.

The second difficulty is related to the accuracy of our approximations. The two particle system in Theorem 8 can be put in a fairly standard slow-fast format. Namely let (q, v) denote the position and velocity of **p** and (Q, V) those of **P**. Put $\varepsilon = 1/\sqrt{M}$ and denote $x = (q, v/\|v\|)$ and $y = (Q, V)$ (note that $\|v\|$ can be recovered from x and y due to the energy conservation). Then x and y transform at the nth collision by

$$\begin{aligned} x_{n+1} &= T_{y_n}(x_n) + \mathcal{O}(\varepsilon), \\ y_{n+1} &= y_n + B(x_n, y_n) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{21}$$

If $T_y(x)$ is a smooth hyperbolic map, the following averaging theorem holds [39]. Let $W \ni (x_0, y_0)$ be a submanifold in the unstable cone, almost parallel to the x-coordinate space (i.e. $y \approx y_0$ on W), and such that $\dim W$ equals the dimension unstable subspace. Then for $|\ln \varepsilon| \ll n \ll 1/\varepsilon$ and any smooth observable A we have

$$\int_W A(x_n, y_n)\, dx_0 = \int A(x, y)\, dm^{y_0}(x) + \varepsilon\, \omega(A, y_0) + o(\varepsilon), \tag{22}$$

where m^{y_0} denotes the SRB measure of the map $T_{y_0}(x)$. This result is a local version of Theorem 7 (consider the case $y_n \equiv y_0$!). In the presence of singularities, however, only a weaker estimate is obtained in [23]:

$$\int_W A(x_n, y_n)\, dx_0 - \int A(x, y)\, dm^{y_0}(x) = \mathcal{O}(\varepsilon\, |\ln \varepsilon|). \tag{23}$$

The extra factor $|\ln \varepsilon|$ appears because we have to wait $\mathcal{O}(|\ln \varepsilon|)$ iterates before the image of W under the unperturbed (billiard) map becomes sufficiently uniformly distributed in the collision space, and at each iteration we have to throw away a subset of measure $\mathcal{O}(\varepsilon)$ in the vicinity of singularities where the shadowing is impossible. The weak estimate (23) was sufficient for time intervals $\mathcal{O}(M^{2/3})$ considered in [23] since the corresponding error term in the expression for V_n, see (17), is

$$\mathcal{O}\left(\frac{n}{M} \times \frac{\ln M}{\sqrt{M}}\right) = \mathcal{O}\left(\frac{\ln M}{M^{5/6}}\right)$$

because $n = \mathcal{O}(M^{2/3})$. This error term is much smaller than the typical value of the velocity, $V_n \sim M^{-2/3}$.

However for $n \sim M$ the above estimate is not good as the error term would far exceeds the velocity itself. To improve the estimate (23) we have to incorporate the vicinity of singularities into our analysis. As the singularities are one-dimensional curves, we expect points falling into their vicinities to have a limit distribution, as

$\varepsilon \to 0$, whose density is smooth on each singularity curve. Finding this distribution requires an accurate counting of billiard orbits passing near singularities. Such counting techniques have been applied to negatively curved manifolds [62], and we hope to extend them to billiards.

Another interesting model involving large mass ratio is so-called piston problem. In that model a container is divided into two compartments by a heavy insulating piston, and these compartments contain particles at different temperatures. If the piston were infinitely heavy, it would not move and the temperature in each compartment would remain constant. However, if the mass of the piston M is finite the temperatures would change slowly due to the energy and momenta exchanges between the particles and the piston. There are several results about infinite particle case (see [21] and references wherein) but the case when the number of particles is finite but grows with M is much more complicated (see [54]). On the other hand if the number of particles is fixed and M tends to infinity then it was shown in [88], [69] under the assumption of ergodicity of billiard in each half of the container that after rescaling time by $1/\sqrt{M}$ the motion of the piston converges to the Hamiltonian system

$$\ddot{Q} = \Delta P := \frac{K^- \ell}{2\pi \operatorname{Area}(\mathcal{D}^-)} - \frac{K^+ \ell}{2\pi \operatorname{Area}(\mathcal{D}^+)}$$

where $\mathcal{D}^-(\mathcal{D}^+)$ is the part of the container to the left(right) of the piston $K^-(K^+)$ is the energy of the particles in $\mathcal{D}^-(\mathcal{D}^+)$ and ℓ is the length of the piston so that ΔP is the pressure difference. In particular if $\Delta P = 0$ and piston is initially at rest then the system does not move significantly during the time $\sqrt{M}$ and the question is what happens on a longer time scale. For the infinite system it was shown in [21] that the motion converges to a diffusion process with the drift in the direction of the hotter gas. In the finite system (for example in a stadium container) this process will be accompanied by simultaneous heating of the piston so that the system may develop rapid ($\dot{Q} \sim \frac{1}{\sqrt{M}}$) oscillations. A similar phenomenon was observed numerically in [25] for a system of $M^{2/3}$ particles in a 3D container. Those oscillations may be responsible for the fact that the system of [25] approaches its thermal equilibrium in $t \sim M^a$ units of time with some $1 < a < 2$ (computer experiments showed that $a \approx 1.7$).

If there is only one particle on either side of the piston the formula (17) suggests that the time of relaxation to equilibrium is of order M, as in $n \sim M$ collisions the heavy disk will reach its maximum velocity $\|V_n\| \sim \sqrt{n}/M = 1/\sqrt{M}$; to prove this we need to improve our approximations along the above lines.

6. Infinite measure systems

Here we discuss several systems with infinite invariant measure, which can serve as tractable models of some non-equilibrium phenomena.

In ergodic theory, systems with infinite (σ-finite) invariant measure are often regarded as exotic and attract little attention. However, hyperbolic and expanding maps with infinite invariant measure appear, more and more often, in various applications. Recently Lenci [55], [56] extended Pesin theory and Sinai's (fundamental) ergodic theorem to unbounded dispersing billiard tables (regions under the graph of a positive monotonically decreasing function $y = f(x)$ for $0 \le x < \infty$), where the collision map, and often the flow as well, have infinite invariant measures.

Another example that we already mentioned is the periodic Lorentz gas with a diffusive particle, but this one can be reduced, because of its symmetries, to a finite measure system by factoring out the $\mathbb{Z}^2$ action (Section 2). The simplest way to destroy the symmetry is to modify the location (or shape) of finitely many scatterers in $\mathbb{R}^2$. We call these finite modifications of periodic Lorentz gases.

Theorem 10. *Consider a periodic Lorentz gas with finite horizon. Then*

(a) ([57]) *its finite modifications are ergodic;*

(b) ([40]) *its finite modifications satisfy Central Limit Theorem with the same covariance matrix as the original periodic gas does.*

The proof of part (a) is surprisingly short. Every finite modification is recurrent, because if it was not, then the particle would not come back to the modified scatterers after some time, so it would move as if in a periodic domain, but every periodic Lorentz gas is recurrent (Theorem 2). Ergodicity then follows by [79].

The proof of (b) uses an analogy with a simple random walk (already observed in Section 2). Recall the proof of Central Limit Theorem for finite modifications of simple random walks [89]. Let ξ_n be a simple random walk on $\mathbb{Z}^2$ whose transition probabilities are modified at one site (the origin). Define $\tilde{\xi}_n$ as follows: initially we set $\tilde{\xi}_0 = \xi_0 = 0$, for every $n \ge 0$ we put

$$\tilde{\xi}_{n+1} - \tilde{\xi}_n = \begin{cases} \xi_{n+1} - \xi_n & \text{if } \xi_n \neq 0, \\ X_n & \text{if } \xi_n = 0, \end{cases}$$

where $X_n = \pm e_i$, $i = 1, 2$, is a random unit step independent of everything else. Then $\tilde{\xi}_n$ is a simple random walk and

$$|\xi_n - \tilde{\xi}_n| \le \mathrm{Card}\{k \le n : \xi_k = 0\}. \tag{24}$$

Since the number of visit to the origin depends only on the behavior of the walk *outside* of the origin the RHS of (24) is $\mathcal{O}(\ln n)$ (see e.g. [33]) so Central Limit Theorem for $\tilde{\xi}_n$ implies Central Limit Theorem for ξ_n.

The analogue of (24) for Lorentz gas if the following. Let $\tilde{\mathbb{B}}_\alpha$ denote the space of α Hölder continuous functions on the collision space of our periodic Lorentz gas with a finite modification, such that every $A \in \tilde{\mathbb{B}}_\alpha$ differs from a periodic function only on a compact set and the periodic part has zero mean. Then if $x_n = (q_n, v_n)$

denotes the position and velocity of the particle after the nth collision and x_0 has a smooth distribution ν with compact support, then for $A \in \tilde{\mathbb{B}}_\alpha$

$$\left|\mathbb{E}(A(x_n))\right| \le c\,\nu\big(\exists k \in [n - c\ln n, n]\colon q_k \text{ is on a modified scatterer}\big) + \mathcal{O}(n^{-100}),$$

where $c > 0$ is a constant. The proof of Theorem 1 (d) given in [90] allows us to estimate the first term here by $\mathcal{O}(\ln^\beta n/n)$ for some $\beta > 0$. The $\ln n$ factor is perhaps an artifact of the proof; on the other hand even for the much simpler case of a modified random walk one has $\mathbb{E}\big(A(\xi_n)\big) \sim c(A)/n$. This implies $\mathbb{E}\big(A(\xi_0)A(\xi_n)\big) \sim \bar{c}(A)/n$. Also there is a quadratic form $q(A)$ such that

$$\mathbb{E}\big(A(\xi_m)A(\xi_{m+n})\big) \sim \bar{q}(A)/(nm), \quad m, n \to \infty. \tag{25}$$

Here we see a new feature of non-stationary systems which does not happen in finite ergodic theory. The correlation series $\sum_{n=1}^{\infty} \mathbb{E}(A(\xi_m)A(\xi_{m+n}))$ diverges for all m but Central Limit Theorem still holds, since the contribution of the off-diagonal terms to $\mathbb{E}(\xi_n^2)$ is much smaller than the contribution of near diagonal terms.

Finite modifications of periodic Lorentz gases are among the simplest billiards with infinite invariant measures, so we hope to move further in their analysis:

Problem 15. Extend (25) to finite modifications of periodic Lorentz gases (with finite horizon).

The reason for this simplicity is that finite modifications are restricted to a 'codimension two' subset of $\mathbb{R}^2$. The particle runs into modified scatterer very rarely, so that its limit distribution is the same as for the unperturbed periodic gas. The situation appears to be different for 'codimension one' modifications.

For example, consider a periodic Lorentz gas and make the particle move in the $N \times N$ box bouncing off its sides and off the scatterers in the box. Denote by $q_N(t)$ the position of the particle at time t.

Problem 16. Prove that $q_N(\tau N^2)/N$ converges, as $N \to \infty$, to the Brownian motion on the unit square with mirror reflections at its boundary.

If the box boundaries are symmetry axis of the Lorentz gas then the result follows easily from Theorem 1 (b) but the general case appears more difficult. In fact if the boundaries of the box are not straight lines (so-called rough boundaries) then one can expect the limit to be different due to trapping and it is an interesting problem to construct such counterexample.

As a more sophisticated example, consider a 'one-dimensional' Lorentz gas – a particle moving in an infinite strip $I = \{(x, y)\colon 0 \le y \le 1\}$ (with identification $(x, 0) = (x, 1)$) and a periodic (in x) configuration of scatterers in I. Suppose a small external force F acts by (8) in a compact domain $x_{\min} \le x \le x_{\max}$. Denote by $q_F(t)$ the position of the moving particle at time t.

Problem 17. Find the limit distribution of $q_F(\tau N)/\sqrt{N}$ as $N \to \infty$.

The analogy with the random walk [89] suggests that $q_F(\tau N)/\sqrt{N}$ should converge to $|\zeta|\eta$ where ζ and η are independent, ζ is a one-dimensional normal distribution $\mathcal{N}(0, \sigma^2)$ where σ^2 is the same as for the system without the field, and η takes values ± 1, so that $\mathbb{P}(\eta = 1) \sim \mathbb{P}(q_n > 0)$ depends on the evolution in the region $x_{\min} \le x \le x_{\max}$. One can further conjecture a functional limit theorem, namely that $q_F(\tau N)/\sqrt{N}$ converges to the so-called skew Brownian motion [49].

While the problems described above could be attacked along the lines of [40], the situation becomes much more difficult if modifications are less regular. In particular very little is known if the location of all scatterers is purely random (if there are infinitely many independent particles in a random Lorentz gas, ergodicity was proven by Sinai [87]).

Problem 18. Do the results of Theorem 10 hold for random Lorentz gas?

The key question is the recurrence of the random Lorentz gas (this issue is irrelevant for infinite particle systems since if one particle wonders to infinity then another one comes to replace it, cf. [31], Chapter 9).

Lenci [58] uses Theorem 10 to show that recurrence holds for an 'open dense set' of Lorentz gases, but this remains to be shown for 'almost every' gas in a measure-theoretic sense.

Problem 17 brings us back to billiards with external forces, see Section 4. We assumed that (8) had an integral of motion. Without this assumption, the system would typically heat up (the particle accelerates indefinitely) or cool down (the particle slows down and stalls). It is interesting to determine which scenario occurs. Denote by $K(t) = \|v(t)\|^2/2$ the particle's kinetic energy at time t.

Problem 19. Consider a Sinai billiards with a velocity-independent external force $\dot{v} = F(q)$. Is $\liminf_{t\to\infty} K(t)$ finite or infinite for most initial conditions?

The particular case of a constant force $F = \text{const}$ is long discussed in physics literature, see [74] and the references therein, but it is yet to be solved mathematically. This model is known in physics as Galton board – a titled plane with a periodic array of pins (scatterers) and a ball rolling on it under a constant (gravitational) force and bouncing off the pins. Due to the conservation of the total energy, the particle accelerates as it goes down the board. Physicists are interested in finding the limit distribution of its position (in a proper time-space scale).

To address this problem observe that if we have a fast particle, i.e. $K(0) = \frac{1}{2\varepsilon}$, then by rescaling the time variable by $s = t/\sqrt{\varepsilon}$ and denoting the rescaled velocity by $u = dx/ds$ we obtain a new equation of motion

$$\frac{du}{ds} = \varepsilon F(q). \tag{26}$$

This system is of type (21) with fast variables $(q, u/\|u\|)$ and a slow variable $T = \|u\|^2/2$. For random Lorentz gases heuristic arguments [74] suggests that in a new

time variable $\tau = \text{const} \cdot \varepsilon^{-2} s$ the limit evolution of T will be given by

$$\dot{T} = \frac{1}{2\sqrt{2T}} + (2T)^{1/4} \dot{w} \tag{27}$$

where $\dot{w}$ is a white noise. The same conclusion is reached in [34] for the geodesic flow on a negatively curved surface in the presence of a weak external force.

As a side remark, observe that the fast motion is obtained here by projecting the right hand side of (26) onto the energy surface, which gives us a thermostated force. In particular (11) plays an important role in the derivation of (27). This shows that the Gaussian thermostat (9), even though regarded as 'artificial' by some physicists, appears naturally in the analysis of weakly forced systems.

We return to the conjecture (27). In terms of our original variables, equation (27) says that $[K(t)]^{3/2}$ is the so-called Bessel square process of index $4/3$, see [75, Chapter XI]. This indicates that $\|v(t)\| \sim t^{1/3}$ so the energy conservation implies $\|q(t) - q(0)\| \sim t^{2/3}$ (cf. [68]). Since the Bessel square process of index $4/3$ is recurrent, it is natural to further conjecture that there is a threshold $K_0 > 0$ such that for almost all initial conditions $\liminf_{t\to\infty} K(t) \le K_0$. This conclusion apparently contradicts a common belief that the particle on the Galton board, see above, generally goes down and accelerates. Rather paradoxically, it will come back up (and hence slow down) infinitely many times! It appears that rigorous mathematics may disagree here with physical intuition, in a spectacular way. The first step in solving this startling paradox would be to extend the averaging theorem (22) to billiards.

References

[1] Aaronson, J., Denker, M., Distributional limits for hyperbolic infinite volume geodesic flows. *Proc. Steklov Inst.* **216** (1997), 174–185.

[2] Aaronson, J., Denker, M., Local limit theorems for partial sums of stationary sequences generated by Gibbs-Markov maps. *Stoch. Dyn.* **1** (2001), 193–237.

[3] Baladi, V., *Positive transfer operators and decay of correlations.* Adv. Ser. Nonlinear Dynam. 16, World Scientific, River Edge, NJ, 2000.

[4] Balint, P., Gouezel, S., Limit theorems in the stadium billiard. *Comm. Math. Phys.* **263** (2006), 461–512.

[5] Bleher, P. M., Statistical properties of two-dimensional periodic Lorentz gas with infinite horizon. *J. Statist. Phys.* **66** (1992), 315–373.

[6] Boca, F., Zaharescu, A., The distribution of the free path in the periodic two-dimensional Lorentz gas in the small-scatterer limit. Manuscript.

[7] Boldrighini C., Bunimovich, L. A., Sinai, Ya. G., On the Boltzmann equation for the Lorentz gas. *J. Statist. Phys.* **32** (1983) 477–501.

[8] Bonetto, F., Daems, D., Lebowitz, J., Properties of stationary nonequilibrium states in the thermostated periodic Lorentz gas I: the one particle system. *J. Statist. Phys.* **101** (2000), 35–60.

[9] Bunimovich, L. A., Billiards that are close to scattering billiards. *Mat. Sb.* **94** (1974), 49–73.

[10] Bunimovich, L. A., On the ergodic properties of nowhere dispersing billiards. *Comm. Math. Phys.* **65** (1979), 295–312.

[11] Bunimovich, L. A., Sinai, Ya. G., Statistical properties of Lorentz gas with periodic configuration of scatterers. *Comm. Math. Phys.* **78** (1980/81), 479–497.

[12] Bunimovich, L. A., Sinai, Ya. G., Chernov, N. I., Statistical properties of two-dimensional hyperbolic billiards. *Russian Math. Surveys* **46** (1991), 47–106.

[13] Buser, P., *Geometry and spectra of compact Riemann surfaces.* Progr. Math. 106, Birkhäuser, Boston 1992.

[14] Burago, D., Ferleger, S., Kononenko, A., Topological entropy of semi-dispersing billiards. *Ergodic Theory Dynam. Systems* **18** (1998), 791–805.

[15] Cercignani, C., Illner, R., Pulvirenti, M., *The mathematical theory of dilute gases*. Appl. Math. Sci. 106, Springer-Verlag, New York 1994.

[16] Chernov, N., Entropy, Lyapunov exponents, and mean free path for billiards. *J. Statist. Phys.* **88** (1997), 1–29.

[17] Chernov, N., Markov approximations and decay of correlations for Anosov flows. *Ann. of Math.* **147** (1998), 269–324.

[18] Chernov, N., Decay of correlations and dispersing billiards. *J. Statist. Phys.* **94** (1999), 513–556.

[19] Chernov, N., Sinai billiards under small external forces. *Ann. Henri Poincaré* **2** (2001), 197–236.

[20] Chernov, N., Advanced statistical properties of dispersing billiards. *J. Statist. Phys.* **122** (2006), 1061–1094.

[21] Chernov, N., On a slow drift of a massive piston in an ideal gas that remains at mechanical equilibrium. *Math. Phys. Electron. J.* **10** (2004), Paper 2, 18 pp.

[22] Chernov, N., in preparation.

[23] Chernov, N., Dolgopyat, D., Brownian Brownian Motion - I. Preprint.

[24] Chernov, N., Eyink, G. L., Lebowitz, J. L., Sinai Ya. G., Steady-state electrical conduction in the periodic Lorentz gas. *Comm. Math. Phys.* **154** (1993), 569–601.

[25] Chernov, N., Lebowitz, J., Dynamics of a massive piston in an ideal gas: Oscillatory motion and approach to equilibrium. *J. Statist. Phys.* **109** (2002), 507–527.

[26] Chernov, N., Markarian, R., *Introduction to the ergodic theory of chaotic billiards.* 2nd ed., IMPA Math. Publ., 24th Brazilian Math. Colloquium, Rio de Janeiro 2003.

[27] Chernov, N., Markarian, R., in preparation.

[28] Chernov, N., Zhang, H.-K., Billiards with polynomial mixing rates. *Nonlinearity* **18** (2005), 1527–1553.

[29] Chernov, N., Zhang, H.-K., A family of chaotic billiards with variable mixing rates. *Stoch. Dyn.* **5** (2005), 535–553.

[30] Conze, J.-P., Sur un critere de recurrence en dimension 2 pour les marches stationnaires, applications. *Ergodic Theory Dynam. Systems* **19** (1999), 1233–1245.

[31] Cornfeld, I. P., Fomin, S. V., Sinai, Ya. G., *Ergodic theory.* Grundlehren Math. Wiss. 245, Springer-Verlag, New York 1982.

[32] Dahlqvist,, P., The Lyapunov exponents in the Sinai billiards in the small scatterer limit. *Nonlinearity* **10** (1997), 159–173.

[33] Darling, D. A., Kac, M., On occupation times for Markoff processes. *Trans. Amer. Math. Soc.* **84** (1957), 444–458.

[34] de la Llave, R., Dolgopyat, D., Stochastic acceleration. In preparation.

[35] Dolgopyat, D., On decay of correlations in Anosov flows. *Ann. of Math.* **147** (1998), 357–390.

[36] Dolgopyat, D., Limit theorems for partially hyperbolic systems. *Trans. Amer. Math. Soc.* **356** (2004), 1637–1689.

[37] Dolgopyat, D., On differentiability of SRB states for partially hyperbolic systems. *Invent. Math.* **155** (2004), 389–449.

[38] Dolgopyat, D., Prelude to a kiss. In *Modern dynamical systems* (ed. by M. Brin , B. Hasselblatt and Ya. Pesin), Cambridge University Press, Cambridge 2004, 313–324.

[39] Dolgopyat, D. Averaging and invariant measures. *Moscow Math. J.*, to appear.

[40] Dolgopyat, D., Szasz, D., Varju, T., Central Limit Theorem for locally perturbed Lorentz process. In preparation.

[41] Donnay, V. J., Using integrability to produce chaos: billiards with positive entropy. *Comm. Math. Phys.* **141** (1991), 225–257.

[42] Erpenbeck, J. J., Wood, W. W., Molecular-dynamics calculations of the velocity-autocorrelation function. Methods, hard-disk results. *Phys. Rev. A* **26** (1982), 1648–1675.

[43] Gallavotti, G., Ornstein, D. S., Billiards and Bernoulli schemes. *Comm. Math. Phys.* **38** (1974), 83–101.

[44] Gaspard, P., Klages, R., Chaotic and fractal properties of deterministic diffusion-reaction processes. *Chaos* **8** (1998), 409–423.

[45] Gilbert, T., Ferguson, C. D., Dorfman, J. R., Field driven thermostated system: a non-linear multi baker map. *Phys. Rev. E* **59** (1999), 364–371.

[46] Golse, F., Wennberg, B., On the distribution of free path lengths for the periodic Lorentz gas-II. *Math. Model. Numer. Anal.* **34** (2000) 1151–1163.

[47] Groeneveld, J., Klages, R., Negative and nonlinear response in an exactly solved dynamical model of particle transport. *J. Statist. Phys.* **109** (2002), 821–861.

[48] *Hard Ball Systems and the Lorentz Gas.* Ed. by D. Szasz, Encyclopaedia Math. Sci. 101, SpringerVerlag, Berlin 2000.

[49] Harrison, J. M., Shepp, L. A., On skew Brownian motion. *Ann. Probab.* **9** (1981), 309–313.

[50] Haydn, N., Vaienti, S., The limiting distribution and error terms for return times of dynamical systems. *Discrete Contin. Dyn. Syst.* **10** (2004), 589–616.

[51] Katok, A., Knieper, G., Pollicott, M., Weiss, H., Differentiability and analyticity of topological entropy for Anosov and geodesic flows. *Invent. Math.* **98** (1989), 581–597.

[52] Kipnis, C., Landim, C., *Scaling limits of interacting particle systems.* Grundlehren Math. Wiss. 320, Springer-Verlag, Berlin 1999.

[53] Lazutkin, V. F., Existence of caustics for the billiard problem in a convex domain. *Math. USSR* (*Izvestya*) **37** (1973), 186–216.

[54] Lebowits, J., Sinai, Ya., Chernov, N., Dynamics of a massive piston immersed in an ideal gas. *Russian Math. Surveys* **57** (2002), 1045–1125.

[55] Lenci, M., Semi-dispersing billiards with an infinite cusp-1. *Comm. Math. Phys.* **230** (2002), 133–180.

[56] Lenci, M., Semidispersing billiards with an infinite cusp-2. *Chaos* **13** (2003), 105–111.

[57] Lenci, M., Aperiodic Lorentz gas: recurrence and ergodicity. *Ergodic Theory Dynam. Systems* **23** (2003), 869–883.

[58] Lenci, M., Typicality of recurrence for Lorentz gases. Preprint.

[59] Liverani, C., On contact Anosov flows. *Ann. of Math.* **159** (2004), 1275–1312.

[60] Liverani, C., Interacting particles. In [48], 179–216.

[61] Machta, J., Power law decay of correlations in a billiard problem. *J. Statist. Phys.* **32** (1983), 555–564.

[62] Margulis, G. A., *On some aspects of the theory of Anosov systems.* With a survey by Richard Sharp: Periodic orbits of hyperbolic flows, Springer Monogr. Math., Springer-Verlag, Berlin 2004.

[63] Markarian, R., Billiards with Pesin region of measure one. *Comm. Math. Phys.* **118** (1988), 87–97.

[64] Markarian, R., Billiards with polynomial decay of correlations. *Ergodic Theory Dynam. Systems* **24** (2004), 177–197.

[65] Melbourne, I., Rapid Decay of Correlations for Nonuniformly Hyperbolic Flows. *Trans. Amer. Math. Soc.*, to appear.

[66] Melbourne, I., Nicol, M., Almost sure invariance principle for nonuniformly hyperbolic systems. *Comm. Math. Phys.* **260** (2005), 131–146.

[67] Melbourne, I., Torok, A., Statistical limit theorems for suspension flows. *Israel J. Math.* **144** (2004), 191–209.

[68] Moran, B., Hoover, W., Bestiale, S., Diffusion in a periodic Lorentz gas. *J. Statist. Phys.* **48** (1987) 709–726.

[69] Neishtadt, A. I., Sinai, Ya. G., Adiabatic piston as a dynamical system. *J. Statist. Phys.* **116** (2004) 815–820.

[70] Olla, S., Varadhan, S. R. S., Yau, H.-T., Hydrodynamical limit for a Hamiltonian system with weak noise. *Comm. Math. Phys.* **155** (1993), 523–560.

[71] Parry, W., Pollicott, M., Zeta functions and the periodic orbit structure of hyperbolic dynamics. *Asterisque* **187-188** (1990).

[72] Pomeau, Y., Resibois, P., Time dependent correlation function and mode-mode coupling theories. *Phys. Rep.* **19** (1975), 63–139.

[73] Ratner, M., The central limit theorem for geodesic flows on n-dimensional manifolds of negative curvature. *Israel J. Math.* **16** (1973), 181–197.

[74] Ravishankar, K., Triolo, L., Diffusive limit of Lorentz model with uniform field from the Markov approximation. *Markov Proc. Rel. Fields* **5** (1999), 385–421.

[75] Revuz, D., Yor, M., *Continuous martingales and Brownian motion.* Grundlehren Math. Wiss. 293, Springer-Verlag, Berlin 1999.

[76] Ruelle, D., Differentiation of SRB states. *Comm. Math. Phys.* **187** (1997), 227–241.

[77] Ruelle, D., Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. *J. Statist. Phys.* **95** (1999), 393–468.

[78] Schmidt, K., On joint recurrence. *C. R. Acad. Sci. Paris Sér. I Math.* **327** (1998), 837–842.

[79] Simanyi, N., Towards a proof of recurrence for the Lorentz process. In *Dynamical systems and ergodic theory* (Warsaw, 1986), ed. by K. Krzyzewski, Banach Center Publ. 23, PWN, Warsaw 1989, 265–276.

[80] Simanyi, N., Proof of the ergodic hypothesis for typical hard ball systems. *Ann. Henri Poincaré* **5** (2004), 203–233.

[81] Simanyi, N., Proof of the Boltzmann-Sinai ergodic hypothesis for typical hard disk systems. *Invent. Math.* **154** (2003), 123–178.

[82] Spohn, H., *Large scale dynamics of interacting particles.* Springer-Verlag, Berlin 1991.

[83] Sinai, Ya. G., On the foundations of the ergodic hypothesis for a dynamical system of statistical mechanics. *Soviet Math. Dokl.* **4** (1963), 1818–1822.

[84] Sinai, Ya. G., Classical dynamic systems with countably-multiple Lebesgue spectrum-II. *Izv. Akad. Nauk SSSR Ser. Mat.* **30** (1966), 15–68.

[85] Sinai, Ya. G., Markov partitions and U-diffeomorphisms. *Funktsional. Anal. i Prilozhen.* **2** (1968), 64–89.

[86] Sinai, Ya. G., Dynamical systems with elastic reflections: Ergodic properties of dispersing billiards. *Russian Math. Surv.* **25** (1970), 137–189.

[87] Sinai, Ya. G., Ergodic properties of a Lorentz gas. *Funktsional. Anal. i Prilozhen.* **13** (1979), 46–59.

[88] Sinai, Ya. G., Dynamics of a massive particle surrounded by a finite number of light particles. *Theor. Math. Phys.* **121** (1999), 1351–1357.

[89] Szasz, D., Telcs, A., Random walk in an inhomogeneous medium with local impurities. *J. Statist. Phys.* **26** (1981), 527–537.

[90] Szasz, D., Varju, T., Local limit theorem for the Lorentz process and its recurrence in the plane. *Ergodic Theory Dynam. Systems* **24** (2004), 257–278.

[91] Szasz, D., Varju, T., Limit laws and recurrence for the planar Lorentz process with infinite horizon. Preprint.

[92] Tsujii, M., Piecewise expanding maps on the plane with singular ergodic properties. *Ergodic Theory Dynam. Systems* **20** (2000), 1851–1857.

[93] Varadhan, S. R. S., Entropy methods in hydrodynamic scaling. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 196–208.

[94] Wojtkowski, M., Invariant families of cones and Lyapunov exponents. *Ergodic Theory Dynam. Systems* **5** (1985), 145–161.

[95] Wojtkowski, M., Principles for the design of billiards with nonvanishing Lyapunov exponents. *Comm. Math. Phys.* **105** (1986), 391–414.

[96] Young, L.-S., Statistical properties of dynamical systems with some hyperbolicity, *Ann. of Math.* **147** (1998), 585–650.

[97] Young, L.-S., Recurrence times and rates of mixing. *Israel J. Math.* **110** (1999), 153–188.

Department of Mathematics, University of Alabama, Birmingham, AL 35294, U.S.A.

E-mail: chernov@math.uab.edu

Department of Mathematics, University of Maryland, College Park, MD 20742, U.S.A.

E-mail: dmitry@math.umd.edu

Some recent progress in geometric methods in the instability problem in Hamiltonian mechanics

Rafael de la Llave

Abstract. We discuss some geometric structures that lead to instability in Hamiltonian systems arbitrarily close to integrable.

The structures covered in this report are joint work with A. Delshams, T. M. Seara and M. Gidea.

Mathematics Subject Classification (2000). Primary 37J40, 37C29, 34C37; Secondary 70H08, 37C50, 34C29.

Keywords. Arnol'd diffusion, normally hyperbolic manifolds, whiskered tori, homoclinic intersections.

1. Introduction

Suppose that we have a system (without friction) and that we perturb it periodically. Will the effect of the forces accumulate or, on the contrary, will the forces average out?

Versions of this question have appeared since ancient times. One of the more ancients versions is the solar system. The Kepler model results from ignoring the interactions between the planets. This model can be solved explicitly and the system persists forever. We can think of the real solar system as a small modification of the Kepler model and wonder if the constant push and pull between the planets will eventually accumulate or whether they would average out and the orbits will remain bounded. See [LP66], [Arn63b] for surveys of the problem at different times. Similar problems appear in many areas of applications. For example, in the study of plasma confinement, or accelerator physics, one can rather easily exhibit confinement for idealized models which ignore mutual interactions, imperfections, etc. One can wonder whether including back the approximations of the model will spoil the confinement.

In other areas of applications, the instabilities are features that have to be sought out and exploited. (Indeed, one of the main goals of human kind has been to get as far as possible with as little effort as possible.) For example, in spacecraft dynamics, there is a great interest in finding ways of moving satellites using the (free!) gravitational forces rather than the (rather expensive!) forces generated by the engines [Bel04]. In theoretical chemistry an important problem is to understand whether the vibrations

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

of one part of a molecule will affect other parts or whether small perturbations will lead to dissociation [JVMU96].

Given the importance and difficulty of the problem, it is no surprise that it has received substantial attention from mathematicians and more applied scientists and that it has been studied by a variety of methods, both rigorous and heuristic.

In this lecture we cannot hope to do justice to this extremely rich history or to survey the recent developments in this very active area. Among several others, we mention the papers [Ber04], [BB02], [BBB03], [Bes96], [BCV01], [BT99], [BK04], [CY04b], [CY04a], [CG94], [CG98], [CP02], [CG03], [Cre03], [Dou88], [DLC83], [FM01], [FM03], [Itu96], [Kal03, LM05], [MS02], [MS04], [Moe96], [Moe02], [Tre02], [Tre04] and the announcements [Xia98], [Mat95], [Mat02]. This list is clearly incomplete!

The only goal of this lecture is to present a concrete point of view, namely, to explain the results in the papers [DdlLS00], [DdlLS03a], [DdlLS03b], [dlL02], [DdlLS05], [GdlL06b] as well as some work in progress (as of Dec. 05) based on the same circle of ideas. We cannot attempt here the much needed systematic survey of the area. We will not even attempt to cover the area of geometric methods and will omit topics such as the *separatrix map*, the study of phenomena that happen in adiabatic phenomena at resonances [Nei86], [NSV03], [IdlLNV02] or the generation of unbounded orbits in Newtonian systems taking advantage of the singularities [Ger91], [Xia92].

Our only goal of this lecture is to present the milestones in a particular approach to the problem. The passage from one milestone to the next is accomplished using a toolkit of standard techniques in the geometric theory of dynamical systems (normal hyperbolicity, averaging methods, KAM theory, Melnikov theory, obstruction theory, and correctly aligned windows). A less standard tool is the scattering map for a normally hyperbolic manifold, which we describe in Section 4.1). Many of the steps could be achieved in different ways. We certainly expect that more techniques will be incorporated in the near future to make the proofs sharper or shorter.

Since we are mainly covering material which is already published or available in much more detailed form, we have omitted important details and assumptions, hence we have not stated theorems. For precise statements and detailed proofs we refer to the references quoted.

2. A mathematical formulation of the instability problem

We will consider a mechanical system. That is a Hamiltonian system defined on a exact symplectic manifold. In some of the models we will discuss, it will be in fact, the product of a torus times and an Euclidean space.[1]

[1] If one considers Hamiltonian systems defined in more general symplectic manifolds, the problems of stability may be very different [Her91].

We will be concerned with situations where our system is close to "integrable." That is, our system can be written as

$$H = H_0(I, \varphi) + \varepsilon H_1(I, \varphi, t) \tag{1}$$

where H_0 is supposed to be integrable and H_1 is periodic (or quasi-periodic) in t.

"Integrable" is often an ill-defined term. In this lecture, we will consider two notions of integrable:

$$H_0 = \sum_{i=1}^{d} h_i(I_i), \tag{2}$$

$$H_0 = \sum_{i=1}^{d} \tilde{h}_i(I_i, \varphi_i). \tag{3}$$

Typical examples are $h_i(I_i) = \frac{1}{2} m_i I_i^2$, $\tilde{h}_i(I_i) = \pm\big(\frac{1}{2m_i} I_i^2 + V_i(\varphi_i)\big)$ with V_i functions with non-degenerate critical points (we will assume $V_i'(0) = 0$, $V_i''(0) > 0$). It is common usage to refer to (2) as "a-priori stable" and (3) as "a-priori unstable", at least when there are hyperbolic fixed points in (3). In the classical terminology used in [AKN88], they are called, respectively, "completely integrable" and "integrable with separable variables". Note that for the systems H_0, the quantities $h_i, \tilde{h}_i$ are conserved quantities. Nevertheless, in (3) the quantities $\tilde{h}_i$ have critical points and indeed, it is impossible to transform them into action-angle variables across the separatrices.

The problem we will consider is to give conditions on H_0 and on the ε-jet of H_1 which guarantee that for $0 < \varepsilon \ll 1$, there are orbits for which some of the variables I experience changes of order 1. This is not the only problem formulated in the literature (see e.g. [Arn68], [Arn04]) but it is the one we will considerer here. The author remembers a round table in S'Agaro [Sim99] in which the organizers asked a distinguished and broad panel to make standard a precise definition of Arnol'd instability. The unanimous consensus was that it was better to let each author make a precise definition of the problems considered in the paper. We note that many papers starting with [HM82] give the name Arnol'd diffusion to the existence of intersections between whiskered tori in (3). This is weaker than the problem we consider here since it ignores the *large gap problem*. See Section 5.

We would also like to describe quantitatively and qualitatively the orbits found. In particular, we would like to describe them in geometric terms and provide geometric features such as estimates on the time they take to move, Hausdorff measure of the orbits described and statistical properties.

Remark. We point out that the distinction between a-priori stable and a-priori unstable models makes sense only for models with one parameter. The model (6) is close to a-priori stable systems, but if $\varepsilon \ll \mu$ we can treat it by methods associated with a-priori unstable systems.

Another important set of models that will be described in Section 4 are systems described by the Hamiltonian

$$\begin{aligned} &H : T^*M \times \mathbb{T}^d \to \mathbb{R}, \\ &H = H_0(p,q) + V(q, \omega t) \end{aligned} \tag{4}$$

where (p, q) denotes a point in T^*M and ω is an external perturbation. We will assume that $H_0(\lambda p, q) = \lambda^2 H_0(p, q)$. Note that the potential term is homogeneous of degree zero.

In the models (4), we would like to find conditions on for which H_0 – which is conserved when $\varepsilon = 0$ – changes by amounts of order 1 whenever $\varepsilon > 0$.

The models (4) were introduced in [Mat95] with H_0 a Riemannian metric and $M = \mathbb{T}^2$, the method presented in [Mat95] is variational. The discussion presented here in section 4 will be based on [DdlLS00], [DdlLS05], [dlL02], [GdlL06b]. Related results are in [BT99]. We note that the geometric methods do not require that the metric is positive definite, so that they apply to Lorenz metrics too. A more detailed comparison between the related results is in [DdlLS05].

Closely related to the models (4) are models the form

$$H(p, q, t) = \frac{1}{2} p^2 + V_n(q) + V_m(q, t) \tag{5}$$

where $p, q \in \mathbb{R}^d$, $d \geq 2$, V_n, V_m are homogeneous of degree n, m respectively, $n > m$, $n > 2$,$V_n > 0$, V_m periodic or quasi-periodic in t. The fact that different terms have different homogeneities makes the geometric analysis similar to that of the models (4). Nevertheless, there are some differences, since the energy surface of (5) has less topology.

The models (5) are extensions to higher dimensions of the models introduced in [Lit66a], [Lit66b] for $d = 1$. A detailed survey of the results on $d = 1$ and important simplifications and corrections to the proofs in the literature is [Lev92] (the original paper [Lit66a], contains a serious error). For $d = 1$, if the terms are sufficiently close to polynomial, [LZ95] show that the orbits remain bounded.

Remark. The main reason to consider periodic or quasi-periodic perturbations is that this is the case that appears in many applications and also as intermediate models after averaging. (Some excellent references for averaging theory are [AKN88], [LM88].)

If one considers more general dependence on time, there is no reason to expect that all orbits average out and indeed, in many cases, it has recently been shown that one should expect instability [Ort04].

2.1. Some early mathematical results on stability. In spite of extremely insightful pioneering works [Poi99], [Lya92] it is not unfair to say that the first definitive and systematic mathematical results to deal with the stability problem arrived in the late 1950s.

The KAM theorem [Kol79], [Arn63a], [Mos66b], [Mos66a] – see also [dlL01] for a modern review – showed that, for a set of initial conditions of large measure, the oscillations of actions remain $O(\varepsilon^{1/2})$ for all time. The applications to celestial mechanics are particularly subtle. (see [Arn63b], [Féj04]) because the system – as many systems of physical interest – fails to satisfy the "generic" assumptions made in many results.

The result [Neh77], [Loc92], [DG96] showed that under "steepness" properties on H_0, one gets stability of $O(\varepsilon^a)$ for very long times $O(\exp(-1/\varepsilon^b))$ for some positive a, b. Hence, in many systems, one can only expect oscillations of the actions of size $\varepsilon^{1/2}$ for all conditions for exponentially long times or for all times in sets of large measure.

It is important to realize that in the a-priori unstable models, the hypotheses of KAM and Nekhoroshev theorem fail in neighborhoods of the separatrices where action-angle coordinates cannot be introduced.

2.2. Some early mathematical results on instability. The first real progress in the mathematical treatment of the problem of instability is the paper [Arn64]. This paper introduces the remarkable example

$$H = \frac{1}{2}p^2 + \frac{1}{2}I^2 + \mu(\cos q - 1) + \varepsilon(\cos q - 1)(\sin\varphi + \cos t), \tag{6}$$

and shows that for $0 < \varepsilon \ll \mu \ll 1$ there are orbits for which the action changes order 1. The mechanism introduced there has served as the basis of much progress. I guess that it will be hard to find a 4 page paper that has generated so much.

The striking example (6) lead to the problems of instability being termed "Arnol'd diffusion". The use of "diffusion" should not imply that it has anything to do with the heat equation.

Some other early results on instability which we will not be able to discuss are [Pus95], [Dou89], [Sit60], [Ale71]. The last two papers consider problems in celestial mechanics and use methods of normally hyperbolic manifolds, so are somewhat related to the methods described here.

2.3. Heuristic studies. Starting in the late 1960s there were extensive numerical studies on the instability problem by many authors. Notably, among many others, we will just mention the surveys [Chi79], [ZZN$^+$89], [TLL80], [Ten82].

Even if these studies were not rigorous, they identified several possible geometric mechanisms for instability and gave empirical mechanisms for their existence. Perhaps the most important fact uncovered by the numerical experiments was that the diffusion is caused by resonances. That is, the trajectories move along the regions where $k \cdot \nabla H_0 = 0$ for $k \in \mathbb{Z}^d$. These curves form the so-called "Arnold web" [Chi79], [ZZN$^+$89], [LR02]. It also became clear that there are different mechanisms at play. A very lucid – albeit non-rigorous – explanation of different mechanisms at play in instability is [Ten82].

3. The example of [Arn64] and the large gap problem

The analysis of [Arn64] is based on a few easy geometric observations, which we will review briefly now in language that we will use later.

We observe that for $0 < \mu$ the manifold $\Lambda = \{A_2 = 0\}$ is invariant. Moreover, Λ is foliated by invariant tori. The stable and unstable manifolds of these tori coincide.

The crucial point of the example is that the ε-dependent term vanishes together with its gradient in Λ. Therefore Λ and the dynamics on it remain unaltered when we make ε positive (but sufficiently small).

Even if the ε term does not have any effect on the manifold Λ it does have an effect on the stable and unstable manifolds. As it turns out, this effect can be computed perturbatively. The idea of the calculation is that the stable and unstable manifolds depend smoothly on parameters. Therefore, the first order term can be computed simply by integrating the variational equations. Some of these calculations in particular cases appear already in [Poi99]. Nowadays these calculations are referred to in the literature as *Melinkov method*.

The last step of the argument of [Arn64] is to show that given any sequence of tori τ_i such that the motion on them is minimal, and such that $W^u_{\tau_i} \pitchfork W^s_{\tau_{i+1}}$ then there is an orbit that follows them.

We note that, because of the exponential convergence, any point which converges to the torus converges to a concrete orbit: $W^{s,u}_{\tau_i} = \bigcup_{x_i \in \tau_i} W^{s,u}_{x_i}$. Therefore, given a sequence of tori, whose manifolds have intersections, there is a sequence of points $x_i \in \tau_i$ such that $W^u_{\tilde{x}_i} \pitchfork W^s_{\tilde{x}_{i+1}}$.

If we start in τ_i, we can wait long enough to move δ-close to $\tilde{x}_i$. Then, we can just move to $W^u_{\tilde{x}_i}$, and then arrive ε-close to τ_{i+1}. Then, a jump allows us to get into τ_{i+1} so that we can get to the next transition point.

If we follow the procedure sketched above, the orbits for each δ are different, so that one needs a separate argument to conclude that there is a true orbit. In the original paper, this is accomplished by a method, which is essentially topological – the obstruction method – but which uses some mild differentiability properties (some version of the λ lemma).

In the words of J. Moser's review in Mathscinet on the 4 page gem is: "The details of the proof must be formidable, although the idea of the proof is clearly outlined". By now, all the details are clearly in print. A modern exposition of the method, including some generalizations is [FM00]. A rather different approach to the analysis of the example in [Arn64] is in [Bes96], which uses variational methods rather than geometric to obtain orbits that shadow the connections between the tori. Unfortunately, we cannot describe the deep and numerous developments generated by [Arn64]. We refer to the comments in [Arn04]. The reader should be warned that many of these papers differ in technical – but crucial! – details. For example, many of the implementations of the obstruction argument on the literature differ on whether or not the method applies to infinitely long orbits. Several important papers assume that one of the terms of the normal form around the torus is zero or that the

stable and unstable bundles are trivial. A sharp version of the argument of [AA67] with references to several other variants is in [DdlLS05]. Other arguments will be described in Sections 7, 9.

3.1. The large gap problem. If rather than taking a perturbation which vanishes on Λ, we had taken a generic perturbation, the tori in Λ that have a rational frequency (resonances) would break, [Poi99], [Tre91], [dlLW04]. These destroyed tori will leave a gap of size $\varepsilon^{1/2}a_i + O(\varepsilon)$ where a_i is a coefficient that depends on the resonance. The fact that outside of these gaps one can find whiskered tori is proved in [Val00].

Since the first order calculations can only predict connections $O(\varepsilon)$ – with a coefficient that is exponentially small in μ – we see that the argument in [Arn64] does not conclude that there are orbits that transverse the gaps generated by a resonance. This is what has been called the "large gap problem". A very lucid discussion of this problem can be found in [Moe96].

This is somewhat unfortunate because the heuristic and numerical explorations suggest that diffusion is more intense near resonance.

Remark. The above discussion has been restricted to second order resonances for flows. That is, $\{k \in \mathbb{Z}^d \mid \frac{\partial H_0}{\partial A} \cdot k = 0\}$ is a 2-dimensional module. One can also wonder about what happens near higher order resonances.

For higher order resonances, the gaps between whiskered tori are larger and the normal forms are not "integrable", so that the discussion of Section 5 does not apply.

Nevertheless, we note that these resonances happen in sets of higher codimension so that using the methods of Section 5, it is possible to construct trajectories that detour around them [DdlLS06a]. A heuristic description of the role of higher order resonances can be found in [Chi79].

4. The role of normally hyperbolic manifolds

One of the main observations of [DdlLS00] (which was crucial for [DdlLS05], [DdlLS03a], [DdlLS03b], [GdlL06b], [CY04b], [CY04a], [Kal03], [BK04] is that the main geometric structure organizing the instability is the presence of a normally hyperbolic invariant manifold Λ whose stable and unstable manifolds intersect transversely.

This invariant manifold Λ acts like a distribution center or a hub. Orbits come to it and get rearranged to exit in a different direction. By returning to Λ again and again, the orbits can change the action systematically. As we will see, in some mechanisms, the orbits gain energy while staying close to Λ and in others, the main gain of energy during the homoclinic excursions.

4.1. The scattering map. A particularly useful tool to keep track of homoclinic excursions to a normally hyperbolic manifold is the scattering map introduced in

[DdlLS00]. Essentially the same idea in a slightly more restrictive context appeared in [Gar00].

Given a family of homoclinic excursions, the scattering map – very similar to the scattering matrix in quantum mechanics – gives the behavior in the distant future as a function of the behavior in the distant past.

The main idea of the scattering map is that the theory of normally hyperbolic invariant manifolds provides a very efficient geometric description of the orbits with a certain asymptotic behavior. By reducing the description of the excursions to the theory of normally hyperbolic manifolds, we obtain a theory which is analytically well behaved. Moreover, the scattering map has very nice geometric properties. By taking the geometry in consideration, it is possible to develop very efficient perturbative calculations [DdlLS06b].

4.1.1. The theory of normally hyperbolic invariant manifolds. We recall some of the results of [Fen72], [Fen74], [HPS77], [Pes04] on normally hyperbolic manifolds. Given a normally hyperbolic manifold Λ, the papers above give an efficient description of the orbits that converge (with a good exponential rate) to Λ, either in the future or in the past. These papers show that these orbits lie on manifolds, which we will denote as W^s_Λ, W^u_Λ, W^s_x, W^u_x. Moreover, these manifolds depend smoothly on parameters. Therefore, by expressing as much as possible the excursions in terms of intersections of invariant manifolds, we obtain remarkable geometric properties and smooth dependence on parameters that will be used to obtain rather explicit perturbative calculations.

Denoting by $W^{s,u}_x$ the set of points whose iterates converge exponentially fast – with an appropriate exponential rate – to the iterates of x,

$$W^s_\Lambda = \bigcup_{x\in\Lambda} W^s_x\,, \quad W^u_\Lambda = \bigcup_{x\in\Lambda} W^s_x \tag{7}$$

and that the decompositions in (7) are a foliation. That is, if $W^s_x \cap W^s_{\tilde x} \neq \emptyset$ then $x = \tilde x$.

The leaves $W^{s,u}_x$ are as smooth as the flow (or map). The map $x \to W^{s,u}_x$ may be significantly less differentiable than the flow (or map) depending on ratios of rates of contraction and expansions. In the applications here, we will have regularities as high as desired since the manifolds have motions close to integrable.

4.1.2. Definition of the scattering map. Let $y \in W^s_\Lambda \cap W^u_\Lambda$ satisfy

$$T_y W^s_\Lambda + T_y W^u_\Lambda = T_y M. \tag{8}$$

By the implicit function theorem, we can find a locally unique manifold $\Gamma \subset W^s_\Lambda \cap W^u_\Lambda$ such that all its points satisfy also (8). Then $T_y(W^s_\Lambda \cap W^u_\Lambda) = T_y\Gamma$.

Given $y \in W^s_\Lambda$, we can associate to it $\Omega_+(y) \in \Lambda$ determined uniquely by $y \in W^s_{\Omega_+(y)}$. Analogously, given $y \in W^u_\Lambda$ we associate to it $\Omega_-(y) \in \Lambda$ determined uniquely by $y \in W^u_{\Omega_-(y)}$.

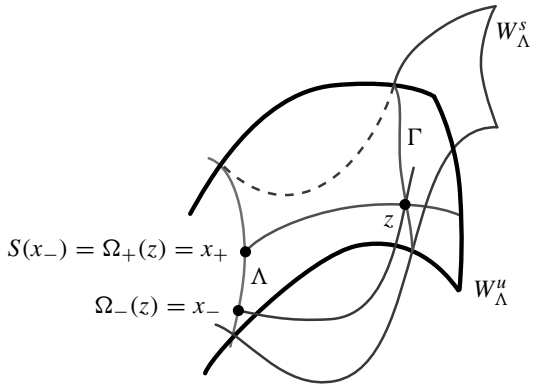


Figure 1. Illustration of the scattering map associated to an intersection Γ.

It is easy to see counting dimensions that, then, the stable and unstable manifolds have a clean intersection along Γ. That is,

$$T_y W^s_{\Omega_+(y)} \oplus T_y\Gamma = T_y W^s_\Lambda; \quad T_y W^u_{\Omega_-(y)} \oplus T_y\Gamma = T_y W^u_\Lambda.$$

Moreover, by taking Γ sufficiently small if necessary, the implicit function theorem ensures that $\Omega_\pm$ are local diffeomorphisms from Γ to their images.

When Ω_- is a local diffeomorphism from Γ to its image we define the mapping S^Γ by

$$S^\Gamma(x_-) = \Omega_+ \circ (\Omega_-)^{-1}. \tag{9}$$

The domain of S^Γ is $H^\Gamma_- \equiv \Omega_-(\Gamma)$. We denote the range of S^Γ as H^Γ_+. We will assume that S^Γ is a diffeomorphism from H^Γ_- to H^Γ_+. By the implicit function theorem, this is true if we have (8) and we take Γ small enough.

Note that, under the assumption that Γ is small enough, that we use to obtain the local uniqueness, it could happen that H^Γ_- is a strict subset of Λ. Nevertheless, for the applications in [DdlLS03b], it is important to observe that the domain of the scattering map can be chosen uniformly for $|\varepsilon|$ small enough.

4.1.3. Geometric properties of the scattering map. Many properties of the scattering map studied in [DdlLS00], [Gar00], [DdlLS03b] are systematized in [DdlLS06b].

Among the properties of the scattering map established in [DdlLS00], [Gar00], [DdlLS03a], [DdlLS03b], [DdlLS05] we mention somewhat informally, ignoring regularity requirements and some subtle points about domains of definition etc. A detailed discussion appears in [DdlLS06b].

- The map S^Γ is (exact) symplectic when f,Γ,Λ are (exact) symplectic diffeomorphisms and manifolds.

- The map S^Γ depends smoothly on parameters when f depends smoothly on parameters.

- There are rather explicit formulas for the derivative of the scattering map with respect to a parameter. These formulas are particularly explicit in the case of symplectic mappings.

- The scattering map associated to an intersection is locally unique. It may happen that when we propagate around a loop, the scattering map has a monodromy.

The scattering map provides a generalization and a more geometric interpretation of the more customary Melinkov theory, which generally assumes that the limiting behavior is of a specific type (e.g., quasi-periodic).

This generality is crucial for the applications in [DdlLS03a], [DdlLS03b] in which transitions to orbits of different topological types are considered. Much more so in [GdlL06a], [GdlL06b] which consider global phenomena.

We also note that the explicit perturbative formulas for the scattering map developed in [DdlLS06b] are always bona fide convergent integrals.

Remark 4.1. The corresponding discussion of the convergence for the Melinkov integrals is rather subtle [Rob96]. Unfortunately, the literature on Melnikov functions is often wrong because it omits a geometric term and the indefinite integrals have a quasi-periodic integrand. The usual explanation that one can take the limit along a subsequence (which one?) is meaningless and, of course, the real reason for this problem is that the argument presented is incorrect. This affects quite a number of papers in the literature. Of course, some of the conclusions may still be true, but they need separate arguments. The reader is alerted to check for this point in the literature.

5. Overcoming the large gap problem by the method of [DdlLS03b]

The work [DdlLS03b] is concerned with one parameter families of the form (2). To simplify the geometric intuition, we will consider f, the time-1 map of the flow.

The first observation is that by appealing to the theory of normally hyperbolic manifolds, in the models (1), there is a normally hyperbolic manifold Λ and that the time-1 map is exact symplectic. This *"inner"* map can be computed to high enough orders in perturbation theory.

A second observation is that, inside Λ, we can perform averaging procedures to high enough order outside of the resonances, so that, outside the resonances, the system can be considered, in an appropriate system of coordinates, as integrable up to order ε^m and m as large as we want. Therefore, outside the resonances, applying the KAM theorem to the averaged system, we can find KAM tori $\varepsilon^{m/2}$ close with m large.

Remark. For the sake of simplicity, the paper [DdlLS03b] includes the hypothesis that the perturbation is a trigonometric polynomial. This allowed to consider uniform several constants appearing in the analysis of the resonances and, therefore, simplified the detailed calculations.

This assumption can be removed by performing a more delicate analysis of the resonances that takes into account that, if the system is differentiable enough, the size of the resonances decreases rapidly and that, for a fixed ε, only a finite number of resonances have width bigger than ε and need to use secondary tori. This is accomplished in [DH06] which contains very detailed analysis of the geometry around resonances. The hypothesis of polynomial perturbations in [DdlLS03b] to conclude topological instability can also be eliminated using topological methods as in [GdlL06a]. See the discussion in Section 7.

As we will see later, the only resonances that play a role in the argument are the resonances that appear during the first averaging procedure and those that appear during the second averaging procedure. These resonances have size $O(\varepsilon^{1/2})$ and $O(\varepsilon)$ respectively.

One key observation for the method of [DdlLS03b] is that, within the resonances, we can find secondary tori – i.e. contractible tori which were not present in the unperturbed system – which dovetail very precisely in the gaps of the foliation of KAM tori. So that, up to precision $O(\varepsilon^{3/2})$ the dynamics on the invariant manifold Λ can be understood as in Figure 2. We emphasize that the invariant sets are, very approximately, the level sets of an averaged energy.

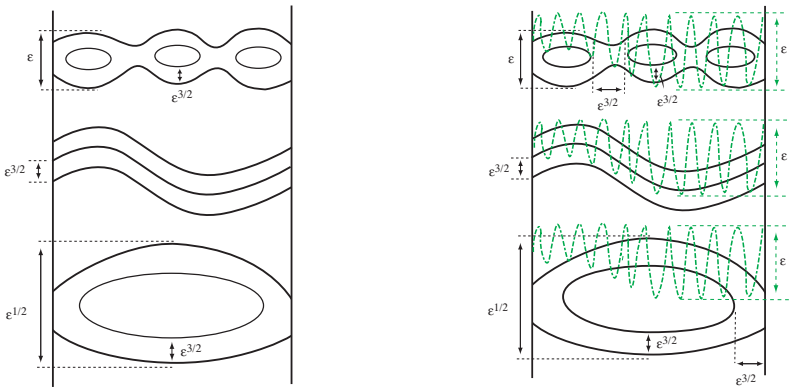


Figure 2. Illustration of the primary and secondary tori in Λ and their images under the scattering map.

If the image under the scattering map of an invariant circle $\tau_1 \subset \Lambda$ crosses transversally another invariant circle $\tau_2 \subset \Lambda$ then $W^u_{\tau_1}$ crosses transversally $W^s_{\tau_2}$.

Lets fix a section in the set of tori (both primary and secondary), say $\theta = \theta_0$ contained in the domain of the scattering map. In [DdlLS03b] there are perturbative calculations of $\frac{d}{d\theta}\bar{H} \circ S|_{\theta=\theta_0}$, where $\bar{H}$ is the averaged energy. If the first order term

happens to be different from zero in an open set (notice it can be chosen independent of ε) then, one can show that given τ_1, it intersects transversally all the tori in a neighborhood of size $O(\varepsilon)$. Using that bounded sets are compact and all the quantities involved are continuous, the order of magnitude estimates have coefficients uniform across a set of size independent of ε. In such a way, one obtains increases of the action in sets of order 1. See a pictorial illustration in Figure 2.

The calculation of the expressions of the angles of intersections is different depending on the types of the tori (primary or secondary) that are involved in the intersections. The hypothesis H5 of [DdlLS03b] is precisely imposing that all the leading terms in the expansion of these angles are different from zero. The geometric meaning of the hypothesis is a transversality condition between the scattering map and the inner dynamics. Since both of them are affected in different ways by the perturbations, it is intuitively clear that the hypothesis should hold for many systems. Given a concrete systems, the conditions can be verified by a finite computation.

Remark. Inside the resonances, one can also find (weakly) hyperbolic periodic orbits for the inner map. Their (weak) stable and unstable manifolds can play the same role as the secondary tori in the construction of transition chains. We refer to [DdlLS03b].

6. Perturbations of geodesic flows and of superlinear oscillators

Variations of the method described in Section 5 can be applied to establish the existence of orbits of unbounded energy in (4) and (5).

In the Mather models (4) we will refer to H_0 as the main term and in the Littlewood models (5) we refer to $\frac{1}{2}p^2 + V_n(q)$ as the main term. The other terms are referred to as perturbative terms.

The reason is that, for high energies, if we scale the time, and the coordinates p, q appropriately, we can map one energy surface of the main term into another. When we perform these scalings, the perturbative terms become small and slow. So that the perturbative parameter ε is just a negative power of the energy. Note that the main term is autonomous and, therefore, the energy is conserved.

The main assumptions (we omit several assumptions on regularity etc. but we refer to the detailed papers) are:

H1 Considering the system generated by the main part restricted to its unit energy surface, there exists a hyperbolic periodic orbit. Its stable and unstable manifolds cross transversally in the unit energy surface.

H2 There are some integrals of the perturbation along the unperturbed orbit that do not vanish identically

H3 The frequency of the external perturbation, ω is Diophantine.

Under these assumptions, the papers [DdlLS00], [DdlLS05] establish the existence of orbits of unbounded energy in (4). The adaptation of these methods to (5) is work in progress. As we will see, there are other methods that allow the elimination of H3, at the price of changing slightly H2.

The idea of the proof is very simple. We observe that, given a hyperbolic orbit in the unit energy, by the scaling properties of the geodesic flow, there are corresponding hyperbolic orbits in any energy surface. The union of all these orbits is a normally hyperbolic manifold for the whole system. If we take the product with the manifold of the phases of the external perturbation, we obtain a normally hyperbolic manifold in the extended system.

In the scaled variables, the external perturbation is slow and small. The fact that the perturbation is slow allows us to conclude that the normally hyperbolic invariant manifold persists, and the fact that the perturbation is slow allows us to average an arbitrarily large number of times. Hence, we can conclude that the invariant manifold contains invariant circles very densely spaced, so that these models do not present the large gap problem and we are in a situation very similar to that of [Arn64].

The calculation of the intersections of the tori can be done rather comfortably using the scattering map. The assumption H2 alluded above is just that the scattering map is non-trivial.

It is quite remarkable that the leading term of the scattering map as the energy grows to infinity is the same expression that was found in [Mat95]. The work [Kal03] uses the geometric approach described above but at some stages of the argument uses variational methods.

It is quite clear that once we choose an orbit and a homoclinic intersection in the main term, the set of lower order terms which verify H2 is a submanifold of infinite codimension.

The assumption of existence of periodic orbits with transversal homoclinic intersections has been studied in great detail for Riemannian manifolds. It holds in great generality. For example, it holds for *all* $C^{2+\alpha}$ Riemannian metrics in a surface of genus greater or equal than 2 and it is known to be generic in many other cases. The paper [Lev97] shows that the hypothesis H1 is verified in the classical Hedlund example.

The paper [BT99] uses instead of periodic orbits the existence of whiskered tori with one dimensional stable and unstable manifolds and with homoclinic intersections. In the case of geodesic flows on surfaces, this coincides with periodic orbits, but in more general cases, both hypothesis are very different.

In the Littlewood models, the hypothesis of the method presented here can be verified when $V_n(q)$ is a small perturbation of $|q_1|^4 + |q_2|^4$, a model for which the conclusions are false.

Remark. Another model that can also be treated by similar methods is billiards with periodically moving boundaries. These models were considered in [KMKPdC96]. In these models, the scaling with the energy is rather different from that in the previous

cases. The motion of the boundary becomes slow but not small, so that, to apply the methods discussed here, one has to add the smallness of the motion as an extra assumption.

Under the hypothesis that the unperturbed billiard has a homoclinic intersection, that the perturbation is small and satisfies a non-degeneracy assumption, using the methods described here, it is possible to show that there are orbits of unbounded energy. Similar results using variational methods have been announced in [Lev05]. It will be quite interesting to develop detailed comparisons between these methods.

7. The method of correctly aligned windows

The method of correctly aligned windows originated in [Eas75], [Eas78], [EM79] and was extended in [ZG04], [GZ04], [GR03], [GR04]. The main tool of the windowing method is the result that shows that if there is a sequence of *well aligned windows*, there is an orbit that follows all of them. As we will see, the construction of such a sequence of well aligned windows follows from an analysis of the geometric structures discussed before. This leads to alternative methods for several steps of the models described in Section 5 which, eliminate some of the hypothesis and provide explicit estimates on the time. They also allow to analyze some other models.

In some ways, the method of well aligned windows can be considered as a topological version of hyperbolicity since it allows shadowing. Nevertheless, an important difference is that the construction of well aligned windows only uses information for a finite number of iterates. As it is well known, hyperbolicity is very delicate to verify since hyperbolicity built up over arbitrarily long times can be destroyed at longer times. Also, one can use windows that are quite extended in some directions. These two advantages are crucial for the applications to diffusion considered in [GdlL06b], [GdlL06a]. Another important advantage is that the fact that well aligned windows are stable under C^0 perturbations. This allows us to use quite comfortably approximately invariant objects.

The following version of the method is taken from [GdlL06b], which in turn relied on [GR03], [GR04].

Definition 7.1. An (n_1,n_2)-window in an n-dimensional manifold M is a compact subset W of M together with a parametrization given by a homeomorphism $\chi^W : [0,1]^{n_1} \times [0,1]^{n_2} \to W$, where $n_1 + n_2 = n$. The set $W^- = \chi^W\left(\partial[0,1]^n_1 \times [0,1]^{n_2}\right)$ is called the 'exit set' and the set $W^+ = \chi^W([0,1]^{n_1} \times \partial[0,1]^{n_2})$ is called the 'entry set' of W. Here ∂ denotes the topological boundary of a set.

Two windows are correctly aligned under some map, provided that the image of the first window under the map can be stretched out and flattened down to a disk crossing the second window all the way through its exist set, so that the induced map on that disk has non-zero degree.

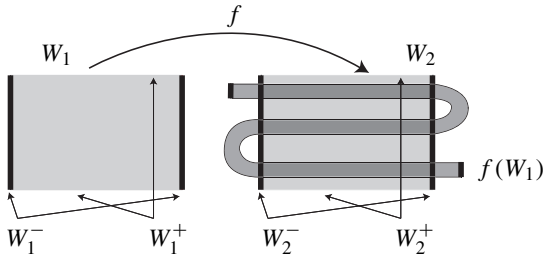


Figure 3. A pair of correctly aligned windows in the plane.

Definition 7.2. Let W_1, W_2 be two (n_1, n_2)-windows in M, and f be a continuous map on M with $f(\mathrm{im}(\chi^{W_1})) \subseteq \mathrm{im}(h^{W_2})$. Denote $f_\chi = (\chi^{W_2})^{-1} \circ f \circ \chi^{W_1}$. We say that W_1 is correctly aligned with W_2 under f provided that the following conditions are satisfied:

(i) $f_\chi((W_1^-)_\chi) \cap (W_2)_\chi = \emptyset, \quad f_\chi((W_1)_\chi) \cap ((W_2)^+_\chi = \emptyset.$

(ii) there exists a point $y_0 \in [0,1]^{n_2}$ such that

(ii.a) $f_\chi([0,1]^{n_1} \times \{y_0\}) \subseteq \mathrm{int}\,([0,1]^{n_1} \times [0,1]^{n_2} \cup (\mathbb{R}^{n_1} \setminus (0,1)^{n_1}) \times \mathbb{R}^{n_2})$,

(ii.b) If $n_1 = 0$, then $f_\chi((W_1)_\chi) \subseteq \mathrm{int}((W_2)_\chi)$. If $n_1 > 0$, then the map $A_{y_0} : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$ defined by $A_{y_0}(x) = \pi_{n_1}\left(f_\chi(x, y_0)\right)$ satisfies

$$A_{y_0}\left(\partial[0,1]^{n_1}\right) \subseteq \mathbb{R}^{n_1} \setminus [0,1]^{n_1}, \quad \deg(A_{y_0}, 0) \neq 0.$$

Here π_x denotes the projection $(x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to x \in \mathbb{R}^{n_1}$.

The main tool of the method is that *"one can see through a sequence of correctly aligned windows"*. Namely, that if we have a sequence of windows $\{W_i\}_{i\in\mathbb{Z}}$, each one of them correctly aligned with the next under the map f, then, there is a point x such that $f^i(x) \in W_i$ for all $i \in \mathbb{Z}$.

In the applications to the models in Sections 5 and 6 considered in [GdlL06b], the windows are constructed as products of intervals in the hyperbolic variables and intervals in the angles and the averaged energy. The windows are laid out around a pseudo-orbit which stays around a constant energy surface but performs jumps at appropriate times.

If we take a window which is an interval slightly offset in the unstable variables and along a homoclinic connection, it will perform a homoclinic excursion and come back as an rectangle very similar to the image under the scattering map of the circle corresponding to a level set of the averaged energy. The transversality between the scattering map and the surfaces of constant averaged energy imply that these windows are well aligned. To construct well aligned windows around the orbits that stay around an invariant circle, the paper [GdlL06b] uses the fact that the inner map has the twist property. We refer to [GdlL06b] for some possible choices of the widths and the choices of exit sets that make the sequence of windows correctly aligned. Compared

with the methods in [DdlLS03b], the method of correctly aligned windows uses the same transversality assumptions but does not need to appeal to the KAM theorem – it suffices to use approximately invariant objects – and it also provides explicit estimates of the time it takes the orbits considered to move order 1 in the action variables.

The paper [GdlL06a] provides with a different choice of windows for the large gap problem (and, quite possibly, different orbits) than those in [GdlL06b]. The main idea is to choose windows which are very wide in the action variables. Indeed, they are wider than the resonances in Λ. As a consequence, the resonances in Λ do not cause any problem in the construction and the paper does not need the transversality hypothesis between the scattering map and the constant average energy foliation. It also can eliminate the hypothesis of the perturbation being a trigonometric polynomial and it suffices just that the problem is differentiable enough. As a further advantage, the orbits constructed move rather fast. The amount it takes them to gain $O(1)$ of energy is $\varepsilon|\log\varepsilon|$, which agrees – up to a multiplicative constant – with lower bounds obtained in [BBB03] for certain models (models without large gaps).

One can hope that the method can be developed much more. Indeed, given the robustness of the windows, it is not necessary to appeal to the theory of normally hyperbolic invariant manifolds. It would suffice to have approximately invariant manifolds which may not be hyperbolic in the strict sense but that expand and contract some directions enough for some finite time. This could be useful in applications since numerical calculations (or fits of dynamical systems obtained from measurements of physical systems) can easily give good information on expansiveness for finite time. Nevertheless, assessing true hyperbolicity is only possible under much more restrictive circumstances.

8. The method of normally hyperbolic laminations

This method is developed in [dlL02]. Some antecedents are the *modulational diffusion* of [Chi79], [Ten82]. On the mathematical side, topological versions in dimension two appear in [Moe02], [EMR01]. For simplicity, we will discuss here only the models (4), but as we will see, they apply to all the models discussed in Section 6.

The main assumptions (again ignoring differentiability assumptions, etc.) are:

H1 There exists a transitive hyperbolic set (e.g. a horseshoe) containing $N \geq 2$ hyperbolic periodic orbits γ_i.

H2 For each of the orbits γ_i above, define $G_i(t) = \frac{1}{|\gamma_i|}\int_0^{|\gamma_i|} \partial_2 V(\gamma_i(s), t)\, ds$. Assume that there exist $0 = a_0 < a_1 < \cdots < a_N = 1$ such that

$$\mathscr{A} \equiv \int_{a_0}^{a_1} G_1(t)\, dt + \int_{a_1}^{a_2} G_2(t)\, dt + \cdots + \int_{a_{N-1}}^{a_N} G_N(t)\, dt \neq 0.$$

Using that $\int_0^1 G_i(t)\, dt = 0$, we will assume without loss of generality that $\mathscr{A} > 0$.

Then, we can find a set of orbits with positive Hausdorff dimension such that for an orbit $x(t)$ in this set,

$$H_0(x(t)) \geq \mathcal{A}t - C.$$

We note that the hypotheses are very abundant. For example, H1 is true for *all* $C^{2+\delta}$ Riemannian metrics in a surface of genus bigger than 1. If we consider a metric of negative curvature (with some pinching conditions in dimensions greater than 2), using results of [GK80] we conclude that H2 is verified for all the potentials which are not of the form $V(q,t) = V_1(q) + V_2(t)$. For potentials of the form above, the conservation of energy shows that there are no orbits of unbounded H_0. For negative curvature metrics, these trivial potentials are the only potentials which fail to have a set of positive Hausdorff dimension of orbits whose energy grows linearly.

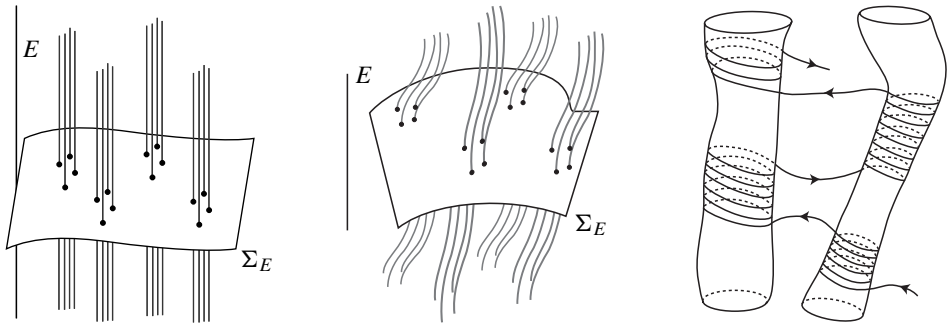


Figure 4. Illustration of the persistence of the invariant laminations and the orbit gaining energy by staying close to periodic orbits.

The idea of the proof is very simple: We observe that, by the scaling properties, for each energy surface there is a transitive hyperbolic set of the geodesic flow. This hyperbolic set satisfies specification and there is a symbolic dynamics that allows us to prescribe times spent in a neighborhood of each of the periodic orbits and connect them by jumps which happen in a scaled time which is uniformly bounded.

If we consider the union of all the hyperbolic sets for all sufficiently large energy, we obtain a normally hyperbolic lamination in the sense of [HPS77].

For high enough energy, the normally hyperbolic lamination persists (one needs to take care of some technicalities such as that the leaves of the foliation have boundary). Indeed, it is possible to find a Hölder map between the new invariant manifold and the old one.

Using the symbolic dynamics given by the normally hyperbolic manifold, we conclude that there are transitions between the periodic orbits happening at times very similar to the a_i in the assumption 2. As can be seen using averaging theory, the meaning of G_i is approximately the gain of energy of orbits that remain close to γ_i but move at high energy. Hence, we can construct orbits which sail along the periodic orbits so that the gains are, on the average $\mathcal{A}$.

Even if the assumptions in Section 6 imply the assumptions in this method, the orbits are very different. Note that the orbits constructed here gain energy near the periodic orbits while the orbits in Section 6 gain energy in the homoclinic excursions.

The method can be adapted to the models (5) and to the billiards with moving boundary. The only differences are that the rate of growths of energy are polynomial and exponential respectively. The adaptation to the time-dependent billiards models requires also the assumption that the motion of the boundary is small. (For these models related results are obtained by variational methods [Lev05].)

9. The scattering map and the obstruction mechanism

By using at the same time arguments similar to the obstruction mechanism of [Arn64], [AA67] it is possible to obtain a mechanism that includes only assumptions on scattering maps. This is work in progress based on preliminary discussions with A. Delshams, M. Gidea and V. Kaloshin. We will only report the simplest case.

We will assume that the manifold Λ is invariant and that it has some homoclinic connections. We will assume without loss of generality that Poincaré's recurrence theorem applies.

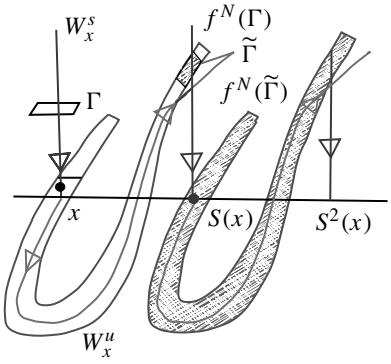


Figure 5. Illustration of the argument in Section 9.

The main inductive lemma starts by considering that Γ is an invariant manifold transverse to W^s_x. It is clear that for all N, we have that $f^N(\Gamma) = \widetilde{\Gamma}$ is transversal to $W^s_{f^N(x)}$. Using the Poincaré recurrence we can assume without loss of generality that $f^N(x)$ comes close to x. On the other hand, by the λ-lemma, the iterates of Γ will be getting aligned with W^u_x. Hence, we can conclude that Γ will intersect transversally the stable manifold of $S(x)$ where S is any of the scattering maps that can be obtained.

In particular, we obtain that if there is a scattering map such that its iterates move order 1, then there is instability in our sense. In the models of the form (1), (3), it suffices that some of the Melnikov integrals do not vanish. Similar results can be obtained for models of the form (4).

10. Conclusions

We have described several different geometric structures that lead to instability in near integrable dynamical systems. These methods rely on normally hyperbolic manifolds acting as a hub for homoclinic excursions. There are other methods, geometric or variational, which we have not covered.

The orbits generated by these different methods have different quantitative properties. It therefore seems that Arnol'd diffusion is not just one phenomenon, but rather a variety of phenomena. It seems that several of the mechanisms are intermingled. When one can find one, one can also find several others.

It seems that the broad array of methods currently under development by a large group of people can lead to answers in different areas. One can hope that even more methods will be brought to bear in a near future.

One can hope that some parts of the very rich heuristic studies – often without a clear statement of conditions of validity – can be clarified by theorems with precise conditions. In particular, it would be quite interesting to develop a rigorous statistical theory of instability.

Relatedly, one can hope that some of the mechanisms can be verified in concrete systems of practical interest or used to design systems with useful properties. From the mathematical point of view, it would also interesting to discuss properties of generic systems.

Acknowledgment. The author has been supported by NSF grants. Stays in Barcelona, crucial for the present work were supported by a visiting professorship of ICREA and by MCyT-FEDER Grant BFM2003-9504.

References

[Ale71] Alexeyev, V. M., Sur l'allure finale du mouvement dans le problème des trois corps. In *Actes du Congrès International des Mathématiciens* (Nice, 1970), Tome 2, Gauthier-Villars, Paris 1971, 893–907.

[AA67] Arnold, V. I., and Avez, A., *Ergodic problems of classical mechanics*. Benjamin, New York 1967.

[AKN88] Arnold, V. I., Kozlov, V. V., and Neishtadt, A. I., *Dynamical Systems III*. Encyclopaedia Math. Sci. 3, Springer-Verlag, Berlin 1988.

[Arn63a] Arnold, V. I., Proof of a theorem of A. N. Kolmogorov on the invariance of quasi-periodic motions under small perturbations. *Russian Math. Surveys* **18** (5) (1963), 9–36.

[Arn63b] Arnol'd, V. I., Small denominators and problems of stability of motion in classical and celestial mechanics. *Russ. Math. Surveys* **18** (1963), 85–192.

[Arn64] Arnold, V. I., Instability of dynamical systems with several degrees of freedom. *Sov. Math. Doklady* **5** (1964), 581–585.

[Arn68] Arnol'd, V. I., A stability problem and ergodic properties of classical dynamical systems. In *Proceedings of the International Congress of Mathematicians* (Moscow, 1966), Izdat. "Mir", Moscow 1968, 387–392.

[Arn04] Arnold, Vladimir I., *Arnold's problems*. Springer-Verlag, Berlin 2004.

[Bel04] Belbruno, Edward, *Capture dynamics and chaotic motions in celestial mechanics*. Princeton University Press, Princeton, NJ, 2004.

[Ber04] Bernard, Patrick, The dynamics of pseudographs in convex Hamiltonian systems. Preprint, 2004; 04-313, http://www.ma.utexas.edu/mp_arc/.

[BB02] Berti, M., and Bolle, P., A functional analysis approach to Arnold diffusion. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **19** (4) (2002), 395–450.

[BBB03] Berti, Massimiliano, Biasco, Luca, and Bolle, Philippe, Drift in phase space: a new variational mechanism with optimal diffusion time. *J. Math. Pures Appl.* (9) **82** (6) (2003), 613–664.

[Bes96] Bessi, Ugo, An approach to Arnol'd's diffusion through the calculus of variations. *Nonlinear Anal.* **26** (6) (1996), 1115–1135.

[BCV01] Bessi, Ugo, Chierchia, Luigi, and Valdinoci, Enrico, Upper bounds on Arnold diffusion times via Mather theory. *J. Math. Pures Appl.* (9) **80** (1) (2001), 105–129.

[BT99] Bolotin, S., and Treschev, D., Unbounded growth of energy in nonautonomous Hamiltonian systems. *Nonlinearity* **12** (2) (1999), 365–388.

[BK04] Bourgain, J., and Kaloshin, V., On diffusion in high-dimensional Hamiltonian systems. *J. Funct. Anal.* **229** (2005), 1–61.

[CY04a] Cheng, Chong-Quin, and Yan, Jun, Arnold diffusion in Hamiltonian systems: 1 a priori unstable case. Preprint, 2004; 04-265, http://www.ma.utexas.edu/mp_arc/.

[CY04b] Cheng, Chong-Quin, and Yan, Jun, Existence of diffusion orbits in a priori unstable Hamiltonian systems. *J. Differential Geom.* **67** (2004), 457–517.

[CG94] Chierchia, L., and Gallavotti, G., Drift and diffusion in phase space. *Ann. Inst. H. Poincaré Phys. Théor.* **60** (1) (1994), 144pp.

[CG98] Chierchia, L., and Gallavotti, G., Erratum drift and diffusion in phase space. *Ann. Inst. H. Poincaré Phys. Théor.* **68** (1998), 135.

[Chi79] Chirikov, B. V., A universal instability of many-dimensional oscillator systems. *Phys. Rep.* **52** (5) (1979), 264–379.

[CP02] Contreras, Gonzalo, and Paternain, Gabriel P., Connecting orbits between static classes for generic Lagrangian systems. *Topology* **41** (4) (2002), 645–666.

[CG03] Cresson, Jacky, and Guillet, Christophe, Periodic orbits and Arnold diffusion. *Discrete Contin. Dyn. Syst.* **9** (2) (2003), 451–470.

[Cre03] Cresson, Jacky, Symbolic dynamics and Arnold diffusion. *J. Differential Equations* **187** (2) (2003), 269–292.

[JVMU96] Diercksen, G. H. F., Von Milczewski, J., and Uzer, T., Computation of the Arnol'd web for the Hydrogen atom in crossed electric and magnetic fields. *Phys. Rev. Lett.* **76** (1996), 2890–2893.

[dlL01] de la Llave, R., A tutorial on KAM theory. In *Smooth ergodic theory and its applications* (Seattle, WA, 1999), ed. by Anatole Katok, Rafael de la Llave, and Yakov Pesin, Amer. Math. Soc., Providence, RI, 2001, 175–292.

[dlL02] de la Llave, R., Orbits of unbounded energy in perturbations of geodesic flows by periodic potentials. A simple construction. Preprint, 2002.

[dlLW04] de la Llave, R., and Wayne, C. E., Whiskered and lower dimensional tori in nearly integrable Hamiltonian systems. *Math. Phys. Electron. J.* **10** (2004), Paper 5, 45 pp. (electronic).

[DdlLS00] Delshams, A., de la Llave, R., and Seara, T. M., A geometric approach to the existence of orbits with unbounded energy in generic periodic perturbations by a potential of generic geodesic flows of $\mathbb{T}^2$. *Comm. Math. Phys.* **209** (2) (2000), 353–392.

[DdlLS03a] Delshams, Amadeu, de la Llave, Rafael, and Seara, Tere M., A geometric mechanism for diffusion in Hamiltonian systems overcoming the large gap problem: announcement of results. *Electron. Res. Announc. Amer. Math. Soc.* **9** (2003), 125–134.

[DdlLS03b] Delshams, Amadeu, de la Llave, Rafael, and Seara, Tere M., A geometric mechanism for diffusion in Hamiltonian systems overcoming the large gap problem: heuristics and rigorous verification on a model. *Mem. Amer. Math. Soc.* **179** (844) (2006), 141pp.

[DdlLS05] Delshams, A., de la Llave, R., and Seara, T. M., Orbits of unbounded energy in generic quasiperiodic perturbations of geodesic flows of certain manifolds. *Adv. Math.* **202** (1) (2006), 64–188.

[DdlLS06a] Delshams, Amadeu, de la Llave, Rafael, and Seara, Tere M., Drift and diffusion in the presence of higher order resonances. Manuscript, 2006.

[DdlLS06b] Delshams, Amadeu, de la Llave, Rafael, and Seara, Tere M., Geometric properties of the scattering map to a normally hyperbolic manifold. Manuscript, 2006.

[DG96] Delshams, Amadeu, and Gutiérrez, P., Effective stability and KAM theory. *J. Differential Equations* **128** (2) (1996), 415–490.

[DH06] Delshams, Amadeu, and Huget, Gemma, The large gap problem in Arnold diffusion for non polynomial perturbations of an apriori unstable Hamiltonian system. Manuscript, 2006.

[DLC83] Douady, Raphaël, and Le Calvez, Patrice, Exemple de point fixe elliptique non topologiquement stable en dimension 4. *C. R. Acad. Sci. Paris Sér. I Math.* **296** (21) (1983), 895–898.

[Dou88] Douady, R., Stabilité ou instabilité des points fixes elliptiques. *Ann. Sci. École Norm. Sup.* (4) **21** (1) (1988), 1–46.

[Dou89] Douady, Raphaël, Systèmes dynamiques non autonomes: démonstration d'un théorème de Pustyl'nikov. *J. Math. Pures Appl.* (9) **68** (3) (1989), 297–317.

[Eas75] Easton, Robert W., Isolating blocks and symbolic dynamics. *J. Differential Equations* **17** (1975), 96–118.

[Eas78] Easton, Robert W., Homoclinic phenomena in Hamiltonian systems with several degrees of freedom. *J. Differential Equations* **29** (2) (1978), 241–252.

[EM79] Easton, Robert W., and McGehee, Richard, Homoclinic phenomena for orbits doubly asymptotic to an invariant three-sphere. *Indiana Univ. Math. J.* **28** (2) (1979), 211–240.

[EMR01] Easton, R. W., Meiss, J. D., and Roberts, G., Drift by coupling to an anti-integrable limit. *Phys. D* **156** (3–4) (2001), 201–218.

[Féj04] Féjoz, Jacques, Démonstration du 'théorème d'Arnold' sur la stabilité du système planétaire (d'après Herman). *Ergodic Theory Dynam. Systems* **24** (5) (2004), 1521–1582.

[Fen72] Fenichel, Neil, Persistence and smoothness of invariant manifolds for flows. *Indiana Univ. Math. J.* **21** (1971/1972), 193–226.

[Fen74] Fenichel, N., Asymptotic stability with rate conditions. *Indiana Univ. Math. J.* **23** (1973/74), 1109–1137.

[FM00] Fontich, E., and Martín, P., Differentiable invariant manifolds for partially hyperbolic tori and a lambda lemma. *Nonlinearity* **13** (5) (2000), 1561–1593.

[FM01] Fontich, Ernest, and Martín, Pau, Arnold diffusion in perturbations of analytic integrable Hamiltonian systems. *Discrete Contin. Dynam. Systems* **7** (1) (2001), 61–84.

[FM03] Fontich, Ernest, and Martín, Pau, Hamiltonian systems with orbits covering densely submanifolds of small codimension. *Nonlinear Anal.* **52** (1) (2003), 315–327.

[Gar00] García, Antonio, Transition tori near an elliptic fixed point. *Discrete Contin. Dynam. Systems* **6** (2) (2000), 381–392.

[Ger91] Gerver, Joseph L., The existence of pseudocollisions in the plane. *J. Differential Equations* **89** (1) (1991), 1–68.

[GdlL06a] Gidea, Marian, and de la Llave, Rafael, Arnold diffusion with optimal time in the large gap problem. Manuscript, 2006.

[GdlL06b] Gidea, Marian, and de la Llave, Rafael, Topological methods in the instability problem of Hamiltonian systems. *Discrete Contin. Dyn. Syst.* **14** (2) (2006), 295–328.

[GR03] Gidea, Marian, and Robinson, Clark, Topologically crossing heteroclinic connections to invariant tori. *J. Differential Equations* **193** (1) (2003), 49–74.

[GR04] Gidea, Marian, and Robinson, Clark, Symbolic dynamics for transition tori-II. In *New advances in celestial mechanics and Hamiltonian systems*, Kluwer/Plenum, New York 2004, 95–108.

[GZ04] Gidea, Marian, and Zgliczyński, Piotr, Covering relations for multidimensional dynamical systems. II. *J. Differential Equations* **202** (1) (2004), 59–80.

[GK80] Guillemin, Victor, and Kazhdan, David, Some inverse spectral results for negatively curved n-manifolds. In *Geometry of the Laplace operator* (Univ. Hawaii, Honolulu, Hawaii, 1979), Proc. Sympos. Pure Math. 36, Amer. Math. Soc., Providence, R.I., 1980, 153–180.

[Her91] Herman, M.-R., Exemples de flots hamiltoniens dont aucune perturbation en topologie C^∞ n'a d'orbites périodiques sur un ouvert de surfaces d'énergies. *C. R. Acad. Sci. Paris Sér. I Math.* **312** (13) (1991), 989–994.

[HPS77] Hirsch, M. W., Pugh, C. C., and Shub, M., *Invariant manifolds*. Lecture Notes in Math. 583, Springer-Verlag, Berlin 1977.

[HM82] Holmes, P. J., and Marsden, J. E., Melnikov's method and Arnol'd diffusion for perturbations of integrable Hamiltonian systems. *J. Math. Phys.* **23** (4) (1982), 669–675.

[IdlLNV02] Itin, A. P., de la Llave, R., Neishtadt, A. I., and Vasiliev, A. A., Transport in a slowly perturbed convective cell flow. *Chaos* **12** (4) (2002), 1043–1053.

[Itu96] Iturriaga, Renato, Minimizing measures for time-dependent Lagrangians. *Proc. London Math. Soc.* (3) **73** (1) (1996), 216–240.

[Kal03] Kaloshin, V., Geometric proofs of Mather's connecting and accelerating theorems. In *Topics in dynamics and ergodic theory*, London Math. Soc. Lecture Note Ser. 310, Cambridge University Press, Cambridge 2003, 81–106.

[KMKPdC96] Koiller, Jair, Markarian, Roberto, Oliffson Kamphorst, Sylvie, and Pinto de Carvalho, Sônia, Static and time-dependent perturbations of the classical elliptical billiard. *J. Statist. Phys.* **83** (1–2) (1996), 127–143.

[Kol79] Kolmogorov, A. N., Preservation of conditionally periodic movements with small change in the Hamilton function. In *Stochastic Behavior in Classical and Quantum Hamiltonian Systems* (Volta Memorial Conf., Como, 1977), Lecture Notes in Phys. 93, Springer-Verlag, Berlin New York 1979, 51–56.

[Lev92] Levi, Mark, On Littlewood's counterexample of unbounded motions in superquadratic potentials. In *Dynamics reported: expositions in dynamical systems*, Springer-Verlag, Berlin 1992, 113–124.

[Lev97] Levi, Mark, Shadowing property of geodesics in Hedlund's metric. *Ergodic Theory Dynam. Systems* **17** (1) (1997), 187–203.

[Lev05] Levi, Mark, Talk at Oberwolfach. 2005.

[LZ95] Levi, M., and Zehnder, E., Boundedness of solutions for quasiperiodic potentials. *SIAM J. Math. Anal.* **26** (5) (1995), 1233–1256.

[Lit66a] Littlewood, J. E., Unbounded solutions of an equation $\ddot{y} + g(y) = p(t)$, with $p(t)$ periodic and bounded, and $g(y)/y \to \infty$ as $y \to \pm\infty$. *J. London Math. Soc.* **41** (1966), 497–507.

[Lit66b] Littlewood, J. E., Unbounded solutions of $\ddot{y} + g(y) = p(t)$. *J. London Math. Soc.* **41** (1966), 491–496.

[LR02] Litvak-Hinenzon, Anna, and Rom-Kedar, Vered, Resonant tori and instabilities in Hamiltonian systems. *Nonlinearity* **15** (4) (2002), 1149–1177.

[LM88] Lochak, P., and Meunier, C., *Multiphase Averaging for Classical Systems*. Appl. Math. Sci. 72, Springer-Verlag, New York 1988.

[LM05] Lochak, Pierre, and Marco, Jean-Pierre, Diffusion times and stability exponents for nearly integrable analytic systems. *Cent. Eur. J. Math.* **3** (3) (2005), 342–397 (electronic).

[Loc92] Lochak, P., Canonical perturbation theory: an approach based on joint approximations. *Uspekhi Mat. Nauk* **47** (6(288)) (1992), 59–140; English translation: *Russian Math. Surveys* **47** (6) (1992), 57–133.

[Lya92] Lyapunov, Aleksander Mikhailovich, *The general problem of the stability of motion*. Taylor and Francis, Bristol 1992.

[MS02] Marco, Jean-Pierre, and Sauzin, David, Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems. *Publ. Math. Inst. Hautes Études Sci.* **96** (2002), 199–275.

[MS04] Marco, Jean-Pierre, and Sauzin, David, Wandering domains and random walks in Gevrey near-integrable systems. *Ergodic Theory Dynam. Systems* **24** (5) (2004), 1619–1666.

[LP66] Marquis de La Place, P. S., *Celestial mechanics*. Vols. I–IV, translated from the French, with a commentary, by Nathaniel Bowditch. Chelsea Publishing Co., Inc., Bronx, N.Y., 1966.

[Mat95] Mather, J. N., Graduate course at Princeton, 95–96, and Lectures at Penn State, Spring 96, Paris, Summer 96, Austin, Fall 96, 1995.

[Mat02] Mather, J. N., Arnold diffusion I: Announcement of results. Preprint, 2002.

[Moe96] Moeckel, Richard, Transition tori in the five-body problem. *J. Differential Equations* **129** (2) (1996), 290–314.

[Moe02] Moeckel, Richard, Generic drift on Cantor sets of annuli. In *Celestial mechanics* (Evanston, IL, 1999), ed. by A. Chenciner, R. Cushman, and C. Robinson, Contemp. Math. 292, Amer. Math. Soc., Providence, RI, 2002, 163–171.

[Mos66a] Moser, J., A rapidly convergent iteration method and non-linear differential equations. II. *Ann. Scuola Norm. Sup. Pisa* (3) **20** (1966), 499–535.

[Mos66b] Moser, J., A rapidly convergent iteration method and non-linear partial differential equations. I. *Ann. Scuola Norm. Sup. Pisa* (3) **20** (1966), 265–315.

[Nei86] Neishtadt, A. I., Change of an adiabatic invariant at a separatrix. *Soviet. J. Plasma Phys.* **12** (1986), 568–573.

[Neh77] Nehorošev, N. N., An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Uspehi Mat. Nauk* **32** (6(198)) (1977), 5–66, 287; English transl. *Russian Math. Surveys* **32** (6) (1977), 1–65.

[NSV03] Neishtadt, Anatoly, Simó, Carles, and Vasiliev, Alexei, Geometric and statistical properties induced by separatrix crossings in volume-preserving systems. *Nonlinearity* **16** (2) (2003), 521–557.

[Ort04] Ortega, R., Unbounded motions in forced newtonian equations. Preprint, 2004.

[Pes04] Pesin, Yakov B., *Lectures on partial hyperbolicity and stable ergodicity*. Zurich Lectures in Advanced Mathematics, European Mathematical Society Publishing House, Zürich, 2004.

[Poi99] Poincaré, H., *Les méthodes nouvelles de la mécanique céleste*, volume 1, 2, 3. Gauthier-Villars, Paris 1892–1899.

[Pus95] Pustyl'nikov, L. D., Poincaré models, rigorous justification of the second law of thermodynamics from mechanics, and the Fermi acceleration mechanism. *Uspekhi Mat. Nauk* **50** (1(301)) (1995), 143–186.

[Rob96] Robinson, Clark, Melnikov method for autonomous Hamiltonians. In *Hamiltonian dynamics and celestial mechanics* (Seattle, WA, 1995), Contemp. Math. 198, Amer. Math. Soc., Providence, RI, 1996, 45–53.

[Sim99] Simó, Carles (editor), *Hamiltonian systems with three or more degrees of freedom*. Kluwer Academic Publishers Group, Dordrecht 1999.

[Sit60] Sitnikov, K., The existence of oscillatory motions in the three-body problems. *Soviet Physics. Dokl.* **5** (1960), 647–650.

[Ten82] Tennyson, Jeffrey, Resonance transport in near-integrable systems with many degrees of freedom. *Phys. D* **5** (1) (1982), 123–135.

[TLL80] Tennyson, J. L., Lieberman, M. A., and Lichtenberg, A. J., Diffusion in near-integrable Hamiltonian systems with three degrees of freedom. In *Nonlinear*

dynamics and the beam-beam interaction (Sympos., Brookhaven Nat. Lab., New York, 1979), ed. by M. Month and J. C. Herrera, Amer. Inst. Physics, New York 1980, 272–301.

[Tre91] Treshchëv, D. V., A mechanism for the destruction of resonance tori in Hamiltonian systems. *Math. USSR-Sb.* **68** (1) (1991), 181–203.

[Tre02] Treschev, D., Trajectories in a neighbourhood of asymptotic surfaces of a priori unstable Hamiltonian systems. *Nonlinearity* **15** (6) (2002), 2033–2052.

[Tre04] Treschev, D., Evolution of slow variables in a priori unstable hamiltonian systems. *Nonlinearity* **17** (5) (2004), 1803–1841.

[Val00] Valdinoci, Enrico, Families of whiskered tori for a-priori stable/unstable Hamiltonian systems and construction of unstable orbits. *Math. Phys. Electron. J.* **6** (2000), Paper 2, 31 pp. (electronic).

[Xia92] Xia, Zhihong, The existence of noncollision singularities in Newtonian systems. *Ann. of Math.* (2) **135** (3) (1992), 411–468.

[Xia98] Xia, Zhihong, Arnold diffusion: a variational construction. In *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), Vol. II, Doc. Math., J. DMV, Extra Vol. ICM Berlin, 1998, 867–877.

[ZZN+89] Zaslavskiĭ, G. M., Zakharov, M. Yu., Neĭshtadt, A. I., Sagdeev, R. Z., Usikov, D. A., and Chernikov, A. A., Multidimensional Hamiltonian chaos. *Zh. Èksper. Teoret. Fiz.* **96** (11) (1989), 1563–1586.

[ZG04] Zgliczyński, Piotr, and Gidea, Marian, Covering relations for multidimensional dynamical systems. *J. Differential Equations* **202** (1) (2004), 32–58.

Department of Mathematics, The University of Texas at Austin and ICES, 1 University Station C1200, Austin, TX 78712-0257, U.S.A.

E-mail: llave@math.utexas.edu

Diagonalizable flows on locally homogeneous spaces and number theory

Manfred Einsiedler and Elon Lindenstrauss*

Abstract. We discuss dynamical properties of actions of diagonalizable groups on locally homogeneous spaces, particularly their invariant measures, and present some number theoretic and spectral applications. Entropy plays a key role in the study of theses invariant measures and in the applications.

Mathematics Subject Classification (2000). 37D40, 37A45, 11J13, 81Q50.

Keywords. Invariant measures, locally homogeneous spaces, Littlewood's conjecture, quantum unique ergodicity, distribution of periodic orbits, ideal classes, entropy.

1. Introduction

Flows on locally homogeneous spaces are a special kind of dynamical systems. The ergodic theory and dynamics of these flows are very rich and interesting, and their study has a long and distinguished history. What is more, this study has found numerous applications throughout mathematics.

The spaces we consider are of the form $\Gamma\backslash G$ where G is a locally compact group and Γ a discrete subgroup of G. Typically one takes G to be either a Lie group, a linear algebraic group over a local field, or a product of such. Any subgroup $H < G$ acts on $\Gamma\backslash G$ and this action is precisely the type of action we will consider here. One of the most important examples which features in numerous number theoretical applications is the space $\mathrm{PGL}(n,\mathbb{Z})\backslash \mathrm{PGL}(n,\mathbb{R})$ which can be identified with the space of lattices in $\mathbb{R}^n$ up to homothety.

Part of the beauty of the subject is that the study of very concrete actions can have meaningful implications. For example, in the late 1980s G. A. Margulis proved the long-standing Oppenheim conjecture by classifying the closed orbits of the group of matrices preserving an indefinite quadratic form in three variables in $\mathrm{PGL}(3,\mathbb{Z})\backslash \mathrm{PGL}(3,\mathbb{R})$ – a concrete action of a three-dimensional group on an eight-dimensional space.

An element h of a linear group G (considered as a group of $n \times n$ matrices over

*The research presented was partially supported by the authors' NSF grants, in particular DMS-0509350 and DMS-0500205. Generous support of the Clay Mathematics Institute of both E. L. and M. E. facilitated much of this research.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

some field K) is said to be *unipotent* if $h - e$ is a nilpotent matrix, e being the identity. Using the adjoint representation one can similarly define unipotent elements for Lie groups. Thanks to the work of M. Ratner, actions of groups H generated by unipotent elements are well understood, and this has numerous applications to many subjects. We refer to [37, Chapter 3], [51] and [59] for more information on this important topic.

In this paper, we focus on the action of diagonalizable groups which of course contain no nontrivial unipotent elements. A prototypical example is the action of the group A of diagonal matrices on $\mathrm{PGL}(n, \mathbb{Z})\backslash \mathrm{PGL}(n, \mathbb{R})$. There is a stark difference between the properties of such actions when $\dim A = 1$ and when $\dim A \geq 2$. In the first case the dynamics is very flexible, and there is a wealth of irregular invariant probability measures and irregular closed invariant sets (though we present some results for one-dimensional actions in § 2.2 under an additional recurrence condition). If $\dim A \geq 2$, the dynamics changes drastically. In particular, it is believed that in this case the invariant probability measures (and similarly the closed invariant sets) are much less abundant and lend themselves to a meaningful classification. Another dynamical property which is less often considered in this context but which we believe is important is the distribution of periodic orbits, i.e. closed orbits of the acting group with finite volume. The purpose of this paper is to present some results in these directions, particularly with regards to the classification of invariant measures, and their applications.

A basic invariant in ergodic theory is the ergodic theoretic entropy introduced by A. Kolmogorov and Ya. Sinai. This invariant plays a surprisingly big role in the study of actions of diagonalizable groups on locally homogeneous spaces as well as in the applications. We discuss entropy and how it naturally arises in several applications in some detail.

In an attempt to whet the reader's appetite, we list below three questions on which the ergodic theoretic properties of diagonalizable flows give at least a partial answer:

- Let $F(x_1, \ldots, x_n)$ be a product of n linear forms in n variables over $\mathbb{R}$. Assume that F is not proportional to such a form with integral coefficients. What can be said about the values F attains on $\mathbb{Z}^n$? In particular, is $\inf_{0 \neq x \in \mathbb{Z}^n} |F(x)| = 0$?
- Let ϕ_i be a sequence of Hecke–Maass cusp forms[1] on $\mathrm{SL}(2, \mathbb{Z})\backslash\mathbb{H}$. What can be said about weak* limits of the measure $|\phi_i|^2\, dm$ (m being the uniform measure on $\mathrm{SL}(2, \mathbb{Z})\backslash\mathbb{H}$)?
- Suppose $n \geq 3$ is fixed. Is it true that any ideal class in a totally real[2] number field K of degree n has a representative of norm $o(\sqrt{\mathrm{disc}(K)})$?

E. L. is scheduled to give a presentation based on this work in the ordinary differential equations and dynamical systems section of the 2006 International Congress of

[1] See §5 for a definition.

[2] A number field K is said to be totally real if any embedding of K to $\mathbb{C}$ is in fact an embedding to $\mathbb{R}$.

Mathematicians. Since much of this is based on our joint work, we have decided to write this paper jointly. Some of the results we present here are joint with P. Michel and A. Venkatesh and will also be discussed in their contribution to these proceedings [50]. We thank H. Oh, M. Ratner, P. Sarnak for comments on our paper. Both M. E. and E. L. have benefited tremendously by collaborations and many helpful discussions related to the topics covered in this survey, and are very thankful to these friends, mentors, colleagues and collaborators.

2. Entropy and classification of invariant measures

2.1. Measures invariant under actions of big diagonalizable groups

2.1.1. We begin by considering a special case where it is widely expected that there should be a complete measure classification theorem for the action of a multidimensional diagonal group. The space we shall consider is $X_n = \mathrm{PGL}(n, \mathbb{Z})\backslash \mathrm{PGL}(n, \mathbb{R})$, which can be identified with the space of lattices in $\mathbb{R}^n$ up to homothety.

Conjecture 2.1. Let A be the group of diagonal matrices in $\mathrm{PGL}(n, \mathbb{R})$, $n \geq 3$. Then any A-invariant and ergodic probability measure μ on the space X_n is homogeneous[3], i.e. is the L-invariant measure on a closed orbit of some group $L \geq A$.

While at present this conjecture remains open, the following partial result is known:

Theorem 2.2 (Einsiedler, Katok, Lindenstrauss [14]). *Let A be the group of diagonal matrices as above and $n \geq 3$. Let μ be an A-invariant and ergodic probability measure on $\mathrm{PGL}(n, \mathbb{Z})\backslash \mathrm{PGL}(n, \mathbb{R})$. If for some $a \in A$ the entropy $h_\mu(a) > 0$ then μ is homogeneous.*

It is possible to explicitly classify the homogeneous measures in this case (see e.g. [42]), and except for measures supported on a single A-orbit none of them is compactly supported. It follows that:

Corollary 2.3. *Let $n \geq 3$. Any compactly supported A-invariant and ergodic probability measure on $\mathrm{PGL}(n, \mathbb{Z})\backslash \mathrm{PGL}(n, \mathbb{R})$ has $h_\mu(a) = 0$ for all $a \in A$.*

For an application of this corollary to simultaneous Diophantine approximation and values of products of linear forms see §4.

2.1.2. We now give a general conjecture, which is an adaptation of conjectures of A. Katok and R. Spatzier [33, Main conjecture] and G. A. Margulis [49, Conjecture 2]. Similar conjectures were made by H. Furstenberg (unpublished).

Let S be a finite set of places for $\mathbb{Q}$ (i.e. a subset of the set of finite primes and ∞). By an S-algebraic group we mean a product $G_S = \prod_{v \in S} G_v$ with each G_v an algebraic

[3]The adjective "algebraic" is also commonly used for this purpose. We follow in this the terminology of [51].

group over $\mathbb{Q}_v$. A $\mathbb{Q}_v$-algebraic group G_v is *reductive* if its unipotent radical is trivial. An S-algebraic group G_S is reductive (semisimple) if each of the G_v is reductive (respectively, semisimple). For any group G and $\Gamma \subset G$ a discrete subgroup, we will denote the image of $g \in G$ under the projection $G \to \Gamma\backslash G$ by $(\!(g)\!)_\Gamma$ or simply $(\!(g)\!)$ if Γ is understood. We shall say that two elements a_1, a_2 of an Abelian topological group A are *independent* if they generate a discrete free Abelian subgroup.

Conjecture 2.4. Let S be a finite set of places for $\mathbb{Q}$, let $G_S = \prod_{v\in S} G_v$ be an S-algebraic group, $G \le G_S$ closed, and $\Gamma < G$ discrete. For each $v \in S$ let $A_v < G_v$ be a maximal $\mathbb{Q}_v$-split torus, and let $A_S = \prod_{v\in S} A_v$. Let A be a closed subgroup of $A_S \cap G$ with at least two independent elements. Let μ be an A-invariant and ergodic probability measure on $\Gamma\backslash G$. Then at least one of the following two possibilities holds:

1. μ is homogeneous, i.e. is the L-invariant measure on a single, finite volume, L-orbit for some closed subgroup $A \le L \le G$.
2. There is some S-algebraic subgroup L_S with $A \le L_S \le G_S$, an element $g \in G$, an algebraic homeomorphism $\phi\colon L_S \to \tilde{L}_S$ onto some S-algebraic group $\tilde{L}_S$, and a closed subgroup $H < \tilde{L}_S$ satisfying $H \ge \phi(\Gamma)$ so that (i) $\mu((L_S \cap G).(\!(g)\!)_\Gamma) = 1$, (ii) $\phi(A)$ does not contain two independent elements, and (iii) the image of μ to $H\backslash\tilde{L}_S$ is not supported on a single point.

Examples due to M. Rees [60] show that μ need not be algebraic, even if $G = \mathrm{SL}(3,\mathbb{R})$ and Γ a uniform lattice; see [12, Section 9] for more details. Such μ arise from algebraic rank one factors of locally homogeneous subspaces as in case 2 of Conjecture 2.4.

2.1.3. We note that the following conjecture is a special case of Conjecture 2.4:

Conjecture 2.5 (Furstenberg). Let μ be a probability measure on $\mathbb{R}/\mathbb{Z}$ invariant and ergodic under the natural action of the multiplicative semigroup $\{p^n q^l\}_{k,l\in\mathbb{Z}^+}$ with p, q multiplicatively independent integers[4]. Then either μ is Lebesgue or it is supported on finitely many rational points.

For simplicity, assume p and q are distinct prime numbers. Then Conjecture 2.5 is equivalent to Conjecture 2.4 applied to the special case of

$$\begin{aligned} A_S &= \{(t_\infty, t_p, t_q) : t_v \in \mathbb{Q}_v^* \text{ for } v \in S\}, \quad S = \{\infty, p, q\},\\ A &= \{(t_\infty, t_p, t_q) \in A_S : |t_\infty| \cdot |t_p|_p \cdot |t_q|_q = 1\},\\ G &= A \ltimes (\mathbb{R} \times \mathbb{Q}_p \times \mathbb{Q}_q),\\ \Gamma &= \Lambda \ltimes \mathbb{Z}[1/pq], \end{aligned}$$

with Λ the group $\{p^k q^l\}_{k,l\in\mathbb{Z}}$ embedded diagonally in A, and $\mathbb{Z}[1/pq]$ embedded diagonally in $\mathbb{R} \times \mathbb{Q}_p \times \mathbb{Q}_q$.

[4] I.e. integers which are not both powers of the same integer.

2.1.4. Following is a theorem towards Conjecture 2.4 generalizing Theorem 2.2. We note that the proof of this more general theorem is substantially more involved.

Theorem 2.6 (Einsiedler, Lindenstrauss [19]). *Let S be a finite set of places for $\mathbb{Q}$ as above, $G_S = \prod_{v\in S} G_v$ a reductive S-algebraic group, and $\Gamma < G_S$ discrete. Let $S' \subset S$ and for any $v \in S'$, let A_v be a maximal $\mathbb{Q}_v$-split torus in G_v. Set $A_{S'} = \prod_{v\in S'} A_v$, and assume $A_{S'}$ has at least two independent elements*[5]*. Let μ be an $A_{S'}$-invariant and the ergodic probability measure on $\Gamma\backslash G_S$. Then at least one of the following two possibilities holds:*

1. *There is some nontrivial semisimple $H < G_S$ normalized by $A_{S'}$ so that* (i) *μ is H invariant,* (ii) *there is some $g \in G_S$ so that $\mu(N_{G_S}(H).(\!(g)\!)_\Gamma) = 1$, and* (iii) *for any $a \in C_{A_{S'}}(H)$, the entropy $h_\mu(a) = 0$.*
2. *There is some $v \in S$ and a reductive $L_v \subset G_v$ of $\mathbb{Q}_v$-rank one satisfying the following: setting $N = C_{G_S}(L_v)$ and $L = NL_v$ (so that $N \lhd L$), there is some $g \in G_S$ so that $\mu(L.(\!(g)\!)_\Gamma) = 1$ and the image of $g^{-1}\Gamma g \cap L$ under the projection $L \to L/N$ is closed.*

Note that option 2 above precisely corresponds to the existence of an algebraic rank one factor of the action as in Conjecture 2.4.

2.2. Recurrence as a substitute for bigger invariance. It is well known that invariance under a one-parameter diagonalizable group is not sufficient to obtain a useful measure classification theorem. On the other hand, it seems that in many situations one can replace additional invariance with a weaker requirement: that of recurrence under some further action.

2.2.1. Let X be a measurable space, equipped with a measure μ, and L a locally compact second countable group acting on X. We first give a definition of recurrence.

Definition 2.7. We say that μ is *recurrent* under L (or L-recurrent) if for every set $B \subset X$ with $\mu(B) > 0$ for a.e. $x \in B$ the set $\{\ell \in L : \ell.x\}$ is unbounded, i.e. has non-compact closure.

This condition is also called conservativity of μ; we find recurrence a more natural term when dealing with the action of general group actions. It can be defined alternatively in terms of *cross-sections*: a set $Y \subset X$ is said to be a cross-section for L if $\mu(L.Y) > 0$ and for every $y \in Y$ there is a neighborhood U of the identity in L so that $\ell.y \notin Y$ for every $\ell \in U \setminus \{e\}$. A cross-section is said to be *complete* if $\mu(X \setminus L.Y) = 0$. We can define recurrence using cross-sections as follows: a measure μ is recurrent under L if there is no cross-section intersecting each L-orbit in at most a single point. This definition is equivalent to the one given in Definition 2.7. An advantage of this viewpoint is that it allows us to consider more refined properties of the action:

[5] Equivalently, that $\operatorname{rank}(A_{S'}) \geq 2$.

Definition 2.8. We say that the L-recurrence of a measure μ is *dominated* by H if there is a complete cross-section $Y \subset X$ for L so that for every $y \in Y$

$$\{\ell \in L : \ell.y \in Y\} \subset H.$$

We say that the L-recurrence of a measure μ is *weakly dominated* by H if there is a (not necessarily complete) cross-section $Y \subset X$ satisfying the same.

2.2.2. We now give a specific rigidity theorem employing recurrence as a substitute for invariance under a multidimensional group. For an application of this theorem to arithmetic quantum unique ergodicity, see §5.

Theorem 2.9 (Einsiedler, Lindenstrauss [20]). *Let v be either ∞ or a finite prime, and let G_v be a semisimple algebraic group over $\mathbb{Q}_v$ with $\mathbb{Q}_v$-rank one. Let A_v be a $\mathbb{Q}_v$-split torus in G_v and let L be an S-algebraic group (S a finite set of places for $\mathbb{Q}$ as above). Let $\Gamma < G_v \times L$ be a discrete subgroup so that $|\Gamma \cap \{e\} \times L| < \infty$. Suppose μ is an A_v-invariant, L-recurrent probability measure on $\Gamma\backslash G_v \times L$, and that for a.e. A_v-ergodic component μ_ξ the entropy $h_{\mu_\xi}(A_v) > 0$. Then a.e. A_v-ergodic component is homogeneous.*

The case $G_v = \mathrm{SL}(2, \mathbb{R})$ was proved in [39] and was used to prove arithmetic quantum unique ergodicity (see §5).

It would be interesting to prove a version of Theorem 2.9 where G_v is replaced by a higher rank algebraic group, e.g. $\mathrm{SL}(3, \mathbb{Q}_v)$, and A_v any algebraic embedding of $\mathbb{Q}_v^* \to G_v$. This seems like a feasible undertaking, but would require new ideas.

2.2.3. It seems desirable to have a general conjecture similar to Conjecture 2.4 where additional invariance is replaced by recurrence. The following seems not completely implausible:

Conjecture 2.10. Let S be a finite set of places for $\mathbb{Q}$, and $G_S = \prod_{v \in S} G_v$ an S-algebraic group and $\Gamma < G_S$ discrete as above. Fix $v \in S$ and let A_v be a rank one $\mathbb{Q}_v$-split torus. Let $L < G_S$ be a closed subgroup commuting with A_v such that $|A_v \cap L| < \infty$. Suppose μ is an A_v-invariant, L-recurrent probability measure. Then there is a $A_v L$-invariant Borel set X' of positive μ-measure so that at least one of the following holds:

1. There are closed subgroups $\tilde{A} \geq A_v$ and $\tilde{L}$ so that (i) $\tilde{A}$ and $\tilde{L}$ commute and have compact intersection[6], (ii) for every $x \in X'$ both $\tilde{L}.x$ and $\tilde{A}.x$ are closed and furthermore $\tilde{A}.x$ has finite $\tilde{A}$-invariant measure, (iii) $\mu|_{X'}$ is $\tilde{A}$-invariant, (iv) the L-recurrence of $\mu|_{X'}$ is dominated by $\tilde{A} \cdot \tilde{L}$.
2. There is a closed subgroup $\tilde{L} < G_S$ commuting with A_v such that for a set of positive μ-measure of $x \in X'$ we have that $\tilde{L}.x$ is closed, and the L-recurrence of $\mu|_{X'}$ is dominated by $\tilde{L}$.

[6]Note that $\tilde{L}$ may be trivial.

If true, this conjecture implies many (if not all) cases of Conjecture 2.4, in particular Conjecture 2.1.

2.2.4. In the notations of the above conjecture, let a be an element in A_v which does not generate a bounded subgroup[7] of A_v. Let

$$G_a^+ = \{g \in G_v : a^{-n} g a^n \to e \quad \text{as } n \to \infty\}$$
$$G_a^- = \{g \in G_v : a^n g a^{-n} \to e \quad \text{as } n \to \infty\}$$
$$G_a^0 = C_{G_v}(a).$$

In the paper [20] we give some nontrivial information on an A_v-invariant, L-recurrent probability measure μ for A_v, L as in the conjecture, under the additional conditions that μ is U_v-recurrent for some $\mathbb{Q}_v$-algebraic subgroup $U_v \leq G_v^-$ such that

1. U_v commutes with L,
2. the U_v recurrence of μ is not weakly dominated by any proper A_v-normalized algebraic subgroup of U_v ,
3. for any $g \in G_v^+$, there is a $u \in U_v$ so that $ugu^{-1} \notin G_v^0 G_v^+$.

The exact conclusions we derive about μ in this case is somewhat technical but in particular they imply Theorem 2.9. Note that one consequence of these conditions is that $h_\mu(A_v) > 0$ (see §3).

2.3. Joinings

2.3.1. Let (X, μ) and (Y, ν) be two measure spaces, and suppose that A is some locally compact group that acts on both (X, μ) and (Y, ν) in a measure preserving way. A *joining* between (X, μ) and (Y, ν) is a measure ρ on $X \times Y$ whose push forward under the obvious projection to X and Y are μ and ν respectively, and which is invariant under the diagonal action of A on $X \times Y$.

One example of a joining which always exists is taking $\rho = \mu \times \nu$ (the *trivial joining*). If $\phi\colon X \to Y$ is a measure preserving map which is A equivariant (i.e. $a.\phi(x) = \phi(a.x)$ for all $a \in A$ and a.e. $x \in X$) then $\rho = (\mathrm{Id} \times \phi)_* \mu$ is a nontrivial joining between (X, μ) and (Y, ν). Note that this joining is supported on the graph of ϕ. Let (Z, η) be another measure space on which A acts preserving the measure. (Z, η) is a *factor*[8] of (X, μ) if there is an A equivariant measurable map $\psi : X \to Z$ so that $\eta = \psi_* \mu$. Any common factor of (X, μ) and (Y, ν) can also be used to give a nontrivial joining called the *relatively independent joining*.

Typically the most interesting case is studying the joinings of a space (X, μ) with itself (called *self joinings*).

[7] I.e. a subgroup with noncompact closure.

[8] It may be more consistent with standard mathematical terminology to call (Z, η) a quotient, but factor is the standard term in ergodic theory.

2.3.2. We now consider joinings between locally homogeneous spaces $\Gamma\backslash G_S$ on which we have an action of a higher rank diagonalizable group. Even though we currently do not have a complete understanding of invariant measures in this context, we are able to give a complete classification of joining between two such actions in many cases.

Theorem 2.11 (Einsiedler, Lindenstrauss [18], [17]). *Let S be a finite set of places for $\mathbb{Q}$, and $G_i = \prod_{v\in S} G_{i,v}$ for $i = 1, 2$ two S-algebraic semisimple groups, $\Gamma_i < G_i$ be lattices*[9]*, and m_i Haar measure on $\Gamma_i\backslash G_i$ normalized to have total mass one. Let $A = \prod_{v\in S'} A_v$ with each A_v a $\mathbb{Q}_v$-split torus and $S' \subseteq S$ satisfying* $\operatorname{rank} A \geq 2$. *Let τ_i $(i = 1, 2)$ be embeddings of A into G_i with the property that*

1. *$\tau_i(A)$ is generated by the subgroups $\tau_i(A) \cap H$ where H runs through the $\mathbb{Q}_v$ simple normal subgroups of $G_{i,v}$ $(v \in S)$.*
2. *For both $i = 1, 2$, there is no S-algebraic group L and an S-algebraic homomorphism $\phi\colon G_i \to L$ so that $\phi(\Gamma_i)$ is discrete and* $\operatorname{rank}(\phi \circ \tau_i(A)) \leq 1$.

Then any ergodic joining between $(\Gamma_1\backslash G_1, m_1)$ and $(\Gamma_2\backslash G_2, m_2)$ is homogeneous[10].

The assumptions of the theorem imply that the action of A on $(\Gamma_i\backslash G_i, m_i)$ $(i = 1, 2)$ is ergodic, and so any joining can be written as the integral of ergodic joinings. The second assumption in Theorem 2.11 regarding the non-existence of rank one factors is clearly necessary. There is no reason to believe the same is true regarding the first assumption.

A special case of the theorem is when G_i are simple algebraic groups over $\mathbb{Q}_v$, v either a finite prime or ∞, and τ_i any algebraic embeddings of $(\mathbb{Q}_v^*)^k$ to G_i, $k \geq 2$. In this case the two assumptions regarding τ_i are automatically satisfied. This case has been treated in [17] using the methods developed by M. E. and A. Katok in [13] (to be precise, only $v = \infty$ is considered in [17] , but there are no difficulties in extending that treatment to $\mathbb{Q}_v$ for any v). The proof of the more general Theorem 2.11 requires also the results in [20].

2.3.3. A different approach to studying joinings was carried out by B. Kalinin and R. Spatzier in [32] using the methods developed by A. Katok and R. Spatzier in [33], [34]. A basic limitation of this technique is that e.g. for actions of split algebraic tori on semisimple or reductive algebraic groups they are able to analyze joinings only if the joining is ergodic not only under the action of the full acting group A but also under the action of certain one parameter subgroups of A. Typically, this is a fairly restrictive assumption, but for joinings which arise from isomorphisms as discussed

[9] I.e. discrete subgroups of finite covolume.

[10] A joining is in particular a measure on $\Gamma_1\backslash G_1 \times \Gamma_2\backslash G_2$ invariant under the diagonal action of the group A. Properties of invariant measures such as ergodicity, homogeneity etc. are in particular equally applicable to joinings.

in § 2.3.1 this assumption is indeed satisfied[11]. This has been used by B. Kalinin and R. Spatzier to classify all measurable isomorphisms between actions of $\mathbb{R}^k$ on $\Gamma_i \backslash G_i$ $(i = 1, 2)$ with G_i a Lie group, $\Gamma_i < G_i$ a lattice and the action of $t \in \mathbb{R}^k$ on $\Gamma_i \backslash G_i$ is given by right translation by $\rho_i(t)$, ρ_i a proper embedding[12] of the group $\mathbb{R}^k$ in G_i whose image is Ad-diagonalizable over $\mathbb{C}$ under some mild conditions on the action of $\rho_i(\mathbb{R}^k)$ on $\Gamma_i \backslash G_i$.

2.3.4. We end our discussion of joinings by noting that extending these joining results to the general context considered in Conjecture 2.4 is likely to be difficult; at the very least it would directly imply Conjecture 2.5.

To see this, let p, q be two multiplicative independent integers, m Lebesgue measure on $\mathbb{R}/\mathbb{Z}$ and μ any other continuous probability measure on $\mathbb{R}/\mathbb{Z}$ invariant and ergodic under the action of the multiplicative semigroup S generated by p and q. Let ρ denote the map $(x, y) \to (x, x + y)$ from $\mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$ to itself. Then since m is weakly mixing for S, the measure $m \times \mu$ is ergodic for S and so $\rho_*(m \times \mu)$ is an ergodic self joining of the action of S on $(\mathbb{R}/\mathbb{Z}, m)$. What we have seen is that a counterexample to Furstenberg's Conjecture 2.5 would give a nonhomogeneous ergodic self joining of $(\mathbb{R}/\mathbb{Z}, m)$, which can be translated to a nonhomogeneous ergodic self joining of the group action considered in § 2.1.3.

One can, however, classify joinings of the actions of commuting endomorphisms of tori with no algebraic projections on which the action degenerates to the action of a virtually cyclic group up to this problem of zero entropy factors. This has been carried out for actions by a group of commuting toral automorphisms satisfying a condition called total nonsymplecticity by B. Kalinin and A. Katok [31] using an adaptation of the methods of A. Katok and R. Spatzier. In [16] the authors deal with general actions of commuting toral automorphisms (without the total nonsymplecticity condition).

2.4. Historical discussion

2.4.1. In 1967 Furstenberg proved that any orbit of the multiplicative semigroup $\{p^k q^l\}_{k,l \in \mathbb{Z}^+}$ on $\mathbb{R}/\mathbb{Z}$ for p, q multiplicatively independent integers is either finite or dense and conjectured Conjecture 2.5 apparently at around the same time. This conjecture seems to have appeared in print only much later (and by other authors quoting Furstenberg). Furstenberg's work was extended to the case of automorphisms of tori and other compact abelian groups by D. Berend (see e.g. [5]).

The first substantial result towards Conjecture 2.5 was published in 1988 by R. Lyons [46], who proved it under the assumption that μ has completely positive entropy for the action generated by the single element p (in particular, μ is ergodic for the single transformation $x \mapsto px \bmod 1$). D. Rudolph [62] showed for p, q relatively prime that it is sufficient to assume that $h_\mu(p) > 0$; this is still the best

[11] The same observation in the context of toral automorphisms was used by A. Katok, S. Katok and K. Schmidt in [36].

[12] I.e. the preimage of compact sets is compact.

result known in this case. The restriction that p, q were relatively prime was lifted by A. Johnson [29]. As Rudolph explicitly pointed out in his paper his proof significantly simplifies if one assumes that μ is ergodic under $x \mapsto px \bmod 1$. Other proofs of this result were given by J. Feldman [24] and B. Host [26] (B. Host actually proves a stronger result that implies Rudolph's). We also note that Host's proof employed recurrence for a certain action which does not preserve the measure, and was one of the motivations for Theorem 2.9.

2.4.2. The first results towards measure classification for actions of diagonalizable groups on quotients of Lie groups and automorphisms of tori were given by A. Katok and R. Spatzier [33], [34]; certain aspects of their work were clarified in [30]. Their proof replaces Rudolph's symbolic description by more geometric concepts, and in particular highlighted the role of conditional measures on invariant foliations on which a subaction acts isometrically. In most cases Katok and Spatzier needed to assume both a condition about entropy and an assumption regarding ergodicity of these subactions. Removing the extra ergodicity assumptions proved to be critical for arithmetic and other applications. M. E. and Katok [12], [13] and E. L. [39] developed two completely different and complementary approaches to proving measure rigidity results in the locally homogeneous context without additional ergodicity assumptions. Both of these techniques were used in [14]. We note that in [39] essential use was made of techniques introduced by M. Ratner to study unipotent flows, particularly her work on rigidity of horocycle flows [56], [54], [55]. Ratner's measure classification theorem [57] and its extensions, such as [58], [48], are also used in these three approaches.

In the context of action by automorphisms on tori no ergodicity assumption was needed by Katok and Spatzier under an assumption they term total nonsymplecticity. A uniform treatment for the general case, using entropy inequalities which should be of independent interest, was given by the authors in [16]. Host [27] has a treatment of some special cases (and even some non commutative actions) by other methods.

2.4.3. We have restricted our attention in this section solely to the measure classification question, but it is interesting to note that already in 1957, J. W. S. Cassels and H. P. F. Swinnerton-Dyer [8] stated a conjecture regarding values of products of three linear forms in three variables (case $n = 3$ of Conjecture 4.1) which is equivalent to Conjecture 4.4 regarding behavior of orbits of the full diagonal group on $\mathrm{SL}(3, \mathbb{Z})\backslash \mathrm{SL}(3, \mathbb{R})$, and which can be derived from Conjecture 2.1. It seems that the first to observe the connection between Furstenberg's work and that of Cassels and Swinnerton-Dyer was G. A. Margulis [47].

3. Brief review of some elements of entropy theory

In this section we give several equivalent definitions of entropy in the context of actions of diagonalizable elements on locally homogeneous spaces, and explain the relations between them.

3.1. General definition of entropy

3.1.1. Let (X, μ) be a probability space. The entropy $H_\mu(\mathcal{P})$ of a finite or countable partition of X into measurable sets measures the average information of $\mathcal{P}$ in the following sense. The partition can be thought of as an experiment or observation whose outcome is the partition element $P \in \mathcal{P}$ the point $x \in X$ belongs to. The information obtained about x from this experiment is naturally measured on a logarithmic scale, i.e. equals $-\log \mu(P)$ for $x \in P \in \mathcal{P}$. Therefore, the average information or *entropy* of $\mathcal{P}$ (with respect to μ) is

$$H_\mu(\mathcal{P}) = -\sum_{P\in\mathcal{P}} \mu(P)\log\mu(P).$$

One basic property of entropy is sub-additivity; the entropy of the refinement $\mathcal{P}\vee\mathcal{Q} = \{P\cap Q : P\in\mathcal{P}, Q\in\mathcal{Q}\}$ satisfies

$$H_\mu(\mathcal{P}\vee\mathcal{Q}) \le H_\mu(\mathcal{P}) + H_\mu(\mathcal{Q}). \tag{3.1}$$

However, this is just a starting point for many more natural identities and properties of entropy, e.g. equality holds in (3.1) if and only if $\mathcal{P}$ and $\mathcal{Q}$ are independent.

The ergodic theoretic entropy $h_\mu(a)$ associated to a measure preserving map $a\colon X \to X$ measures the average amount of information one needs to keep track of iterates of a. To be more precise we need to start with a fixed partition $\mathcal{P}$ (either finite or countable with $H_\mu(\mathcal{P}) < \infty$) and then take the limit

$$h_\mu(a,\mathcal{P}) = \lim_{N\to\infty} \frac{1}{N} H_\mu\Big(\bigvee_{n=0}^{N-1} a^{-n}\mathcal{P}\Big).$$

To get independence of the choice of $\mathcal{P}$ the ergodic theoretic entropy is defined by

$$h_\mu(a) = \sup_{\mathcal{P}:H_\mu(\mathcal{P})<\infty} h_\mu(a,\mathcal{P}).$$

The ergodic theoretic entropy was introduced by A. Kolmogorov and Ya. Sinai and is often called the Kolmogorov–Sinai entropy; it is also somewhat confusingly called the metric entropy even though X often has the additional structure of a metric space and in that case there is a *different* (though related) notion of entropy, the topological entropy (see §4.2.2), which is defined using the metric on X.

3.1.2. Entropy has many nice properties and is manifest in many different ways. We mention a few which will be relevant in the sequel.

A partition $\mathcal{P}$ is said to be a *generating partition* for a and μ if the σ-algebra $\bigvee_{n=-\infty}^{\infty} a^{-n}\mathcal{P}$ (i.e. the σ-algebra generated by the sets $\{a^n.P : n\in\mathbb{Z}, P\in\mathcal{P}\}$) separates points, that is for μ-almost every x, its atom with respect to this σ-algebra is $\{x\}$.[13]

[13] Recall that the atom of x with respect to a countably generated σ-algebra $\mathcal{A}$ is the intersection of all $B\in\mathcal{A}$ containing x and is denoted by $[x]_{\mathcal{A}}$.

The Kolmogorov–Sinai theorem asserts the nonobvious fact that $h_\mu(a) = h_\mu(a, \mathcal{P})$ whenever $\mathcal{P}$ is a generating partition.

Entropy is most meaningful when μ is ergodic. In this case, positive entropy $h_\mu(a) > 0$ means that the entropy of the repeated experiment grows linearly, i.e. every new iteration of it reveals some new information of the point. In fact, one can go to the limit here and say that the experiment reveals new information even when one already knows the outcome of the experiment in the infinite past. Similarly, zero entropy means that the observations in the past completely determine the present one. If μ is an a-invariant but not necessarily ergodic measure, with an ergodic decomposition $\mu = \int \mu_x^{\mathcal{E}}\, d\mu(x)$,[14] then

$$h_\mu(a) = \int h_{\mu_x^{\mathcal{E}}}(a)\, d\mu(x), \tag{3.2}$$

i.e. the entropy of a measure is the average of the entropy of its ergodic components.

3.2. Entropy on locally homogeneous spaces

3.2.1. Let $G = \prod_{v\in S} G_v$ be an S-algebraic group, $\Gamma < G$ discrete, and set $X = \Gamma\backslash G$. The Lie algebra of G_S can be defined as the product of the Lie algebra of the G_v, and the group G acts on its Lie algebra of G by conjugation; this action is called the adjoint representation and for every $a \in G$ the corresponding Lie algebra endomorphisms is denoted by $\operatorname{Ad} a$. Fix an $a \in G$ for which $\operatorname{Ad} a$ restricted to the Lie algebra of each G_v is diagonalizable over $\mathbb{Q}_v$. We implicitly identify between $a \in G$ and the corresponding map $x \mapsto a.x$ from X to itself.

The purpose of this subsection is to explain how the entropy $h_\mu(a)$ of an a-invariant measure μ relates to more geometric properties of X. A good reference for more advanced results along this direction is [48, Section 9] which contains an adaptation of results of Y. Pesin, F. Ledrappier, L. S. Young and others to the locally homogeneous context.

3.2.2. Fundamental to the dynamics of a are the stable and unstable horospherical subgroups G_a^- and G_a^+ introduced in § 2.2.4. Both G_a^- and G_a^+ are unipotent algebraic groups and the Lie algebras of G_a^- (resp. G_a^+) are precisely the sums of the eigenspaces of the adjoint Ad_a of eigenvalue with absolute value less than (resp. bigger than) one. For any $x \in X$ the orbits $G_a^-.x$ and $G_a^+.x$ are precisely the stable and unstable manifolds of x. We will also need the group

$$G_a^0 = \{g \in G : \text{the set } \{a^n g a^{-n}, n \in \mathbb{Z}\} \text{ is bounded}\}.$$

Under our assumptions G_a^0 can be shown to be an algebraic subgroup of G, and if 1 is the only eigenvalue of $\operatorname{Ad} a$ of absolute value one, $G_a^0 = C_G(a)$. This subgroup G_a^0 together with G_a^- and G_a^+ give a local coordinate system of G, i.e. there are

[14] This decomposition has the property that $\mu_x^{\mathcal{E}}$ is ergodic and for a.e. x, the ergodic averages of a function f along the orbit of x converge to $\int f\, d\mu_x^{\mathcal{E}}$.

neighborhoods $V^- \subset G_a^-$, $V^+ \subset G_a^+$, and $V^0 \subset G_a^0$ of e for which $V^+V^-V^0$ is a neighborhood of e in G and the map from $V^+ \times V^- \times V^0$ to $V^+V^-V^0$ is a bijection.

3.2.3. Suppose X is compact. If $\mathcal{P}$ is a finite partition with elements of small enough diameter, then the atoms of x with respect to $\mathcal{A} = \bigvee_{n=1}^{\infty} a^{-n}\mathcal{P}$ is a subset of $V^-V^0.x$ for all $x \in X$ as x and y are in the same $\mathcal{A}$-atom if and only if $a^n.x$ and $a^n.y$ are in the same partition element of $\mathcal{P}$ for $n = 1, 2, \ldots$. In particular, they must be close-by throughout their future, which can only be if the V^+-component of their relative displacement is trivial. Similarly, the atom of x with respect to $\bigvee_{-\infty}^{\infty} a^{-n}\mathcal{P}$ is a subset of $V^0.x$ for all $x \in X$.

Thus even though $\mathcal{P}$ might not be a generator, the atoms of the σ-algebra generated by $a^{-n}\mathcal{P}$ for $n \in \mathbb{Z}$ are small, with each atom contained in a uniformly bounded subset of a single G_a^0-orbit. G_a^0 possesses a metric invariant under conjugation by a, and this implies that a acts isometrically on these pieces of G_a^0-orbits. Such isometric extensions cannot produce additional entropy, and indeed a modification of the proof of the Kolmogorov–Sinai theorem gives that

$$h_\mu(a, \mathcal{P}) = h_\mu(a) \quad \text{for } \textit{all } a\text{-invariant measures } \mu. \tag{3.3}$$

In the non-compact case there is a somewhat weaker statement of this general form that is still sufficient for most applications.

3.2.4. Positive entropy can be characterized via geometric tubes as follows. Let B be a fixed open neighborhood of $e \in G$, and define $B_n = \bigcap_{k=-n}^{n} a^k B$. A tube around $x \in X = \Gamma\backslash G$ is a set of the form $B_n.x$ for some n. Then for a as in § 3.2.1, it can be shown using the Shannon–McMillan–Breiman theorem that if B is sufficiently small, for any measure μ with ergodic decomposition $\int \mu_x^{\mathcal{E}}\, d\mu(x)$

$$h_{\mu_x^{\mathcal{E}}}(a) = \lim_{n\to\infty} \frac{-\log \mu(B_n.x)}{2n} \qquad \text{for } \mu\text{-a.e. } x. \tag{3.4}$$

In particular, positive entropy of μ is equivalent to the exponential decay of the measure of tubes around a set of points which has positive μ-measure, and positive entropy of all ergodic components of μ is equivalent to the same holding for a conull subset of X. We note that (3.4) is a variant of a more general result of Y. Brin and A. Katok [7].

3.2.5. Positive entropy can also be characterized via recurrence. For μ, a as above the following are equivalent: (i) $h_{\mu_x^{\mathcal{E}}}(a) > 0$ for a.e. a-ergodic component $\mu_x^{\mathcal{E}}$, (ii) μ is G_a^--recurrent, and (iii) μ is G_a^+-recurrent (see e.g. [39, Theorem 7.6]).

3.2.6. A quite general phenomenon is upper semi-continuity of entropy $h_\mu(a)$ as a function of μ in the weak* topology[15]. Deep results of Yomdin, Newhouse and Buzzi

[15] A sequence of measures μ_i converges in the weak* topology to μ if for every compactly supported continuous function f one has that $\int f\, d\mu_i \to \int f\, d\mu$.

establish this for general C^∞ diffeomorphisms of compact manifolds, but in our context establishing such semi-continuity is elementary, particularly in the compact case. For non-compact quotients $X = \Gamma\backslash G$ and a sequence of a-invariant probability measures we might have escape of mass in the sense that a weak* limit might not be a probability measure. In that case entropy might get lost (even if some mass remains). However, if we assume that the weak* limit is again a probability measure then upper semi-continuity still holds in this context. The case where the whole sequence is supported on a compact set is discussed in [14, Cor. 9.3]. The general case follows along similar lines, the key step is showing that there is a finite partition capturing almost all of the ergodic theoretic entropy uniformly for the sequence, cf. §3.2.3.

4. Entropy and the set of values obtained by products of linear forms

4.1. Statements of conjectures and results regarding products of linear forms

4.1.1. In this section we consider the following conjecture:

Conjecture 4.1. Let $F(x_1, x_2, \dots, x_n)$ be a product of n linear forms in n variables over the real numbers with $n \geq 3$, and assume that F is not proportional to a homogeneous polynomial with integer coefficients. Then

$$\inf_{0\neq x\in\mathbb{Z}^d} |F(x)| = 0. \tag{4.1}$$

We are not sure what is the proper attribution of this conjecture, but the case $n = 3$ was stated by J. W. S. Cassels and H. P. F. Swinnerton-Dyer in 1955 [8]. In that same paper, Cassels and Swinnerton-Dyer show that Conjecture 4.1 implies the following conjecture of Littlewood:

Conjecture 4.2 (Littlewood (c. 1930)). For any $\alpha, \beta \in \mathbb{R}$,

$$\liminf_{n\to\infty} n \, \|n\alpha\| \, \|n\beta\| = 0,$$

where for any $x \in \mathbb{R}$ we denote $\|x\| = \min_{n\in\mathbb{Z}} |x - n|$.

We let $\mathcal{F}_n$ denote the set of products of n linear forms in n-variables, considered as a subvariety of the space of degree n homogeneous polynomials in n-variables, and $P\mathcal{F}_n$ the corresponding projective variety with proportional forms identified. For any $F \in \mathcal{F}_n$ we let $[F]$ denote the corresponding point in $P\mathcal{F}_n$.

4.1.2. The purpose of this section is to explain how measure classification results (specifically, Corollary 2.3) can be used to prove the following towards the above two conjectures:

Theorem 4.3 (Einsiedler, Katok, Lindenstrauss [14]). 1. *The set of products* $[F] \in P\mathcal{F}_n$ *for which* $\inf_{0\neq x\in\mathbb{Z}^n} |F(x)| > 0$ *has Hausdorff dimension zero.*

2. *The set of* $(\alpha, \beta) \in \mathbb{R}^2$ *for which* $\liminf_{n\to\infty} n \, \|n\alpha\| \, \|n\beta\| > 0$ *has Hausdorff dimension zero.*

Even though Conjecture 4.1 implies Conjecture 4.2, part 2 of Theorem 4.3 does not seem to be a formal consequence of part 1 of that theorem. The proofs, however, are very similar. Related results by M. E. and D. Kleinbock in the S-arithmetic context can be found in [15].

4.1.3. As noted by G. A. Margulis [47], Conjecture 4.1 is equivalent to the following conjecture regarding the orbits of the diagonal group on $X_n = \mathrm{PGL}(n, \mathbb{Z})\backslash \mathrm{PGL}(n, \mathbb{R})$ (cf. Conjecture 2.1 above):

Conjecture 4.4. Let A be the group of diagonal matrices in $\mathrm{PGL}(n, \mathbb{R})$, and X_n as above. Then for any $x \in X_n$ its orbit under A is either periodic (closed of finite volume) or unbounded[16].

The equivalence between Conjecture 4.1 and Conjecture 4.4 is a consequence of the following simple proposition (the proof is omitted):

Proposition 4.5. *The product of n linear forms*

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} (\ell_{i1}x_1 + \cdots + \ell_{in}x_n)$$

satisfies $\inf_{0\neq x\in\mathbb{Z}^n} |F(x)| \geq \delta$ *if and only if there is no* $(\!(g)\!) \in A.(\!(\ell)\!)$ *where* $\ell = (\ell_{ij})_{i,j=1}^n$ *and a nonzero* $x \in \mathbb{Z}^n$ *so that*

$$\|xg\|_\infty^n < \det(g)\delta.$$

In particular, by Mahler's compactness criterion, (4.1) holds if and only if $A.(\!(\ell)\!)$ is unbounded.

The map $F \mapsto N_G(A).(\!(\ell)\!)$, $N_G(A)$ being the normalizer of A in $G = \mathrm{PGL}(n, \mathbb{R})$ gives a bijection between $\mathrm{PGL}(n, \mathbb{Z})$ orbits in $P\mathcal{F}_n$ and orbits of $N_G(A)$ in X_n. Note that $N_G(A)$ is equal to the semidirect product of A with the group of $n \times n$ permutation matrices.

4.1.4. In a somewhat different direction, G. Tomanov [71] proved that if F is a product of n linear forms in n variables, $n \geq 3$, and if the set of values $F(\mathbb{Z}^n)$ is discrete, then F is proportional to a polynomial with integer coefficients. Translated to dynamics, this statement reduces to a classification of all *closed* A-orbits (bounded or unbounded) which was carried out by Tomanov and B. Weiss in [72].

[16] I.e. does not have compact closure.

4.2. Topological entropy and A-invariant closed subsets of X_n

4.2.1. Corollary 2.3 regarding A-invariant measures on X_n which have positive ergodic theoretic entropy with respect to some $a \in A$ implies the following purely topological result towards Conjecture 4.4 (this theorem is essentially [14, Theorem 11.2]):

Theorem 4.6. *Let Y be a compact A-invariant subset of X_n. Then for every $a \in A$ the topological entropy $h_{\mathrm{top}}(Y, a) = 0$.*

Theorem 4.3 can be derived from Theorem 4.6 by a relatively straightforward argument that in particular uses Proposition 4.5 to translate between the orbits of A and Diophantine properties of products of linear forms (and a similar variant to relate orbits of a semigroup of A with the behavior of $\liminf_{k\to\infty} k \|k\alpha\| \|k\beta\|$).

4.2.2. We recall the definition of topological entropy, which is the topological dynamical analog of the ergodic theoretic entropy discussed in §3.1: Let (Y, d) be a compact metric space and $a\colon Y \to Y$ a continuous map[17]. Two points $y, y' \in Y$ are said to be k, ε-*separated* if for some $0 \le \ell < k$ we have that $d(a^\ell.y, a^\ell.y') \ge \varepsilon$. Set $N(Y, a, k, \varepsilon)$ to be the maximal cardinality of a k, ε-separated subset of Y. Then the topological entropy of (Y, a) is defined by

$$H(Y, a, \varepsilon) = \liminf_{k\to\infty} \frac{\log N(Y, a, k, \varepsilon)}{k},$$
$$h_{\mathrm{top}}(Y, a) = \lim_{\varepsilon\to 0} H(Y, a, \varepsilon).$$

We note that in analogy to §3.2.3, for the systems we are considering, i.e. $a \in A$ acting on a compact $Y \subset X_n$ there is some $\varepsilon(Y)$ so that

$$h_{\mathrm{top}}(Y, a) = H(Y, a, \varepsilon) \quad \text{for } \varepsilon < \varepsilon(Y).$$

4.2.3. Topological entropy and the ergodic theoretic entropy are related by the *variational principle* (see e.g. [25, Theorem 17.6] or [35, Theorem 4.5.3])

Proposition 4.7. *Let Y be a compact metric space and $a\colon Y \to Y$ continuous. Then*

$$h_{\mathrm{top}}(Y, a) = \sup_{\mu} h_\mu(a)$$

where the sup *runs over all a-invariant probability measures supported on Y.*

Note that when $\mu \mapsto h_\mu(a)$ is upper semicontinuous (see §3.2.6), in particular if $a \in A$ (identified with the corresponding translation on X_n) and $Y \subset X_n$ compact, this supremum is actually attained by some a-invariant measure on Y.

[17] For Y which is only locally compact, one can extend a to a map $\tilde{a}$ on its one-point compactification $\tilde{Y}$ fixing ∞ and define $h_{\mathrm{top}}(Y, a) = h_{\mathrm{top}}(\tilde{Y}, \tilde{a})$.

4.2.4. We can now explain how Theorem 4.6 can be deduced from the measure classification results quoted in §2

Proof of Theorem 4.6 *assuming Corollary* 2.3. Let $Y \subset X_n$ be compact, and $a \in A$ be such that $h_{\text{top}}(Y, a) > 0$. Then by Proposition 4.7 there is an a-invariant probability measure μ supported on Y with $h_\mu(a) > 0$. Let

$$S_r = \{a \in A : \|a\|,\ \|a^{-1}\| < e^r\}.$$

Since A is a commutative group

$$\mu_r = \int_{S_r} (a'.\mu)\, dm_A(a')$$

(where m_A is Haar measure on A) is also a-invariant, and in addition it follows directly from the definition of entropy that $h_{a'.\mu}(a) = h_\mu(a)$ for any $a' \in A$. Using (3.2) it follows that $h_{\mu_r}(a) = h_\mu(a)$.

Let ν be any weak* limit point of μ_r. Then by semicontinuity of entropy,

$$h_\nu(a) \geq \liminf_{r\to\infty} h_{\mu_r}(a) > 0$$

and since S_r is a Folner sequence in A we have that ν is A-invariant. Finally, since Y is A-invariant, and μ is supported on Y, so is ν. But by Corollary 2.3, ν cannot be compactly supported – a contradiction. □

5. Entropy and arithmetic quantum unique ergodicity

Entropy plays a crucial role also in a completely different problem: arithmetic quantum unique ergodicity. Arithmetic quantum unique ergodicity is an equidistribution question, but unlike most equidistribution questions it is not about equidistribution of points but about equidistribution of eigenfunction of the Laplacian. A more detailed discussion of this topic can be found in [40] and the surveys [63], [65] as well as the original research papers, e.g. [61], [39].

5.1. The quantum unique ergodicity conjecture

5.1.1. Let M be a complete Riemannian manifold with finite volume which we initially assume to be compact. Then since M is compact, $L^2(M)$ is spanned by the eigenfunctions of the Laplacian Δ on M. Let ϕ_n be a complete orthonormal sequence of eigenfunctions of Δ ordered by eigenvalue. These can be interpreted for example as the steady states for Schrödinger's equation

$$-i\frac{\partial \psi}{\partial t} = \Delta\psi$$

describing the quantum mechanical motion of a free (spinless) particle of unit mass on M (with the units chosen so that $\hbar = 1$). According to Bohr's interpretation

of quantum mechanics $\tilde{\mu}_n(A) := \int_A |\phi_n(x)|^2 \, dm_M(x)$ is the probability of finding a particle in the state ϕ_n inside the set A at any given time, m_M denoting the Riemannian measure on M, normalized so that $m_M(M) = 1$. A. I. Šnirel′man, Y. Colin de Verdière and S. Zelditch [69], [9], [75] have shown that whenever the geodesic flow on M is ergodic, for example if M has negative curvature, there is a subsequence n_k of density one on which $\tilde{\mu}_{n_k}$ converge in the weak* topology to m, i.e. $\tilde{\mu}_n$ become equidistributed *on average*. Z. Rudnick and P. Sarnak conjectured that in fact if M has negative sectional curvature, $\tilde{\mu}_n$ become equidistributed *individually*, i.e. that $\tilde{\mu}_n$ converge in the weak* topology to the uniform measure m_M.

5.1.2. As shown in [69], [9], [75] any weak* limit $\tilde{\mu}$ of a subsequence of the $\tilde{\mu}_i$ is the projection of a measure μ on the unit tangent bundle SM of M invariant under the geodesic flow. This measure μ can be explicitly constructed directly from the ϕ_i. We shall call μ the *microlocal lift* of $\tilde{\mu}$. We shall call any measure μ on SM arising in this way a *quantum limit*. A stronger form of Rudnick and Sarnak's conjecture regarding $\tilde{\mu}_n$ is the following (also due to Rudnick and Sarnak).

Conjecture 5.1 (Quantum unique ergodicity conjecture [61]). Let M be a compact Riemannian manifold with negative sectional curvature. Then the uniform measure m_{SM} on SM is the only quantum limit.

There is numerical evidence towards this conjecture in the analogous case of 2D concave billiards by A. Barnett [3], and some theoretical evidence is given in the next two subsections, but whether this conjecture should hold for general negatively curved manifolds remains unclear.

5.2. Arithmetic quantum unique ergodicity

5.2.1. Consider now the special case of $M = \Gamma\backslash\mathbb{H}$, for Γ one of the following:

1. Γ is a congruence sublattice of $\mathrm{PGL}(2, \mathbb{Z})$.
2. D is a quaternion division algebra over $\mathbb{Q}$, split over $\mathbb{R}$ (i.e. $D(\mathbb{R}) := D \otimes \mathbb{R} \cong M(2, \mathbb{R})$). Let $\mathcal{O}$ be an Eichler order in D.[18] Then the norm one elements in $\mathcal{O}$ are a co-compact lattice in $D(\mathbb{R})^*/\mathbb{R}^*$. Let Γ be the image of this lattice under the isomorphism $D(\mathbb{R})^*/\mathbb{R}^* \cong \mathrm{PGL}(2, \mathbb{R})$.

We shall call such lattices *lattices of congruence type* over $\mathbb{Q}$.

5.2.2. If Γ is as in case 1 of §5.2.1, then $M = \Gamma\backslash\mathbb{H}$ has finite volume, but is not compact. A generic hyperbolic surface of finite volume is expected to have only finitely many eigenfunctions of Laplacian in L^2; consequently Conjecture 5.1 needs some modification to remain meaningful in this case. However, a special property

[18] A subring $\mathcal{O} < D$ is said to be an order in D if $1 \in \mathcal{O}$ and every for $\beta \in \mathcal{O}$ its trace and its norm are in $\mathbb{Z}$. An order $\mathcal{O}$ is an Eichler order if it is the intersection of two maximal orders.

of congruence sublattices of $\mathrm{PGL}(2,\mathbb{Z})$ congruence is the abundance of cuspidal eigenfunctions of the Laplacian (in particular, the existence of many eigenfunctions in L^2) on the corresponding surface, and so Conjecture 5.1 as stated is both meaningful and interesting for these surfaces. The abundance of cuspidal eigenfunctions follows from Selberg's trace formula [66]; see [41] for an elementary treatment.

We remark that for congruence sublattices of $\mathrm{PGL}(2,\mathbb{Z})$ the continuous spectrum of the Laplacian is given by Eisenstein series; equidistribution (appropriately interpreted) of these Eisenstein series has been established by W. Luo and P. Sarnak [44] and by D. Jakobson [28].

5.2.3. The lattices given in §5.2.1 have the property that for all but finitely many primes p, there is a lattice Λ_p in $\mathrm{PGL}(2,\mathbb{R}) \times \mathrm{PGL}(2,\mathbb{Q}_p)$ so that

$$\Gamma \backslash \mathrm{PGL}(2,\mathbb{R}) \cong \Lambda_p \backslash \mathrm{PGL}(2,\mathbb{R}) \times \mathrm{PGL}(2,\mathbb{Q}_p)/K_p \tag{5.1}$$

with $K_p = \mathrm{PGL}(2,\mathbb{Z}_p) < \mathrm{PGL}(2,\mathbb{Q}_p)$. For example, for $\Gamma = \mathrm{PGL}(2,\mathbb{Z})$ one can take $\Lambda_p = \mathrm{PGL}(2,\mathbb{Z}[1/p])$ (embedded diagonally in $\mathrm{PGL}(2,\mathbb{R}) \times \mathrm{PGL}(2,\mathbb{Q}_p)$).

5.2.4. The isomorphism (5.1) gives us a map $\pi_p \colon \Lambda_p\backslash\mathbb{H} \times \mathrm{PGL}(2,\mathbb{Q}_p) \to \Gamma\backslash\mathbb{H}$. The group $\mathrm{PGL}(2,\mathbb{Q}_p)$ acts on $\Lambda_p\backslash\mathbb{H} \times \mathrm{PGL}(2,\mathbb{Q}_p)$ by right translation, and using this action we set for every $x \in \Gamma\backslash\mathbb{H}$

$$T_p(x) = \pi_p\left(\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}.\pi_p^{-1}(x)\right);$$

this is a set of $p+1$ points called the *p-Hecke correspondence*. Using this correspondence we define the *p-Hecke operator* (also denoted T_p) on functions on $\Gamma\backslash\mathbb{H}$ by

$$[T_p(f)](x) = p^{-1/2} \sum_{y \in T_p(x)} f(y).$$

The Hecke operators (considered as operators on $L^2(M)$) are self adjoint operators that commute with each other and with the Laplacian, so one can always find an orthonormal basis of the subspace of $L^2(M)$ which corresponds to the discrete part of the spectrum consisting of such joint eigenfunctions. Furthermore, if the spectrum is simple (as is conjectured e.g. for $\mathrm{PGL}(2,\mathbb{Z})$), eigenfunctions of the Laplacian are automatically eigenfunctions of all Hecke operators. These joint eigenfunctions of the Laplacian and all Hecke operators are called *Hecke–Maass cusp forms*.

5.2.5. We define an *arithmetic quantum limit* to be a measure μ on SM which is a quantum limit constructed from a sequence of Hecke–Maass cusp forms (see §5.1.2). The *arithmetic quantum unique ergodicity* question, also raised by Rudnick and Sarnak in [61] is whether the uniform measure on SM is the only arithmetic quantum limit. Some partial results towards answering this question were given

in [61], [38], [74], [64]. Assuming the Riemann hypothesis for suitable automorphic L-functions, T. Watson [73] has shown that the only arithmetic quantum limit for both types of lattices considered in §5.2.1 is the normalized volume measure. In fact, to obtain this conclusion one does not need the full force of the Riemann hypothesis but only subconvexity estimates on the value of these L-functions at $1/2$, which are known for some families of L-functions but not for the ones appearing in Watson's work. Assuming the full force of the Riemann hypothesis gives a rate of convergence of the $\tilde{\mu}_k$ to the uniform measure that is known to be best possible [45].

Using measure classification techniques one can unconditionally prove the following:

Theorem 5.2 (Lindenstrauss, [39]). *Let M be $\Gamma\backslash\mathbb{H}$ for Γ one of the lattice as listed in §5.2.1. Then if M is compact the arithmetic quantum limit is the uniform measure m_{SM} on SM (normalized to be a probability measure). In the noncompact case, any arithmetic quantum limit is of the form cm_{SM} some $c \in [0, 1]$.*

It is desirable to prove unconditionally that even in the non-compact case m_{SM} is the only arithmetic quantum limit. A weaker version would be to prove unconditionally that for any sequence of Hecke–Maass cusp forms ϕ_i and $f, g \in C_c(M)$ with $g \geq 0$

$$\frac{\int f(x)\,|\phi_i(x)|^2\,dm(x)}{\int g(x)\,|\phi_i(x)|^2\,dm(x)} \to \frac{\int f\,dm(x)}{\int g\,dm(x)}.$$

5.2.6. We briefly explain how measure rigidity results are used to prove Theorem 5.2. Let ϕ_i be a sequence of Hecke–Maass cusp forms on $\Gamma\backslash\mathbb{H}$, and let μ be the associated arithmetic quantum limit, which we recall is a measure on SM which is essentially $\Gamma\backslash\mathrm{PGL}(2,\mathbb{R})$. Using the isomorphism (5.1) and in the notations of §5.2.3, one can identify μ with a right K_p-invariant measure, say μ', on $\Lambda_p\backslash\mathrm{PGL}(2,\mathbb{R}) \times \mathrm{PGL}(2,\mathbb{Q}_p)$. Using the fact that ϕ_i is an eigenfunction of the Hecke operators and some elementary fact regarding the regular representations of $\mathrm{PGL}(2,\mathbb{Q}_p)$ one can show that μ' is recurrent under $\mathrm{PGL}(2,\mathbb{Q}_p)$ (see Definition 2.7).

Building upon an idea of Rudnick and Sarnak from [61], J. Bourgain and E. L. [6] have shown that any ergodic component $\mu_x^{\mathcal{E}}$ of μ with respect to the action of the group $a(t) = \left(\begin{smallmatrix} e^t & 0 \\ 0 & e^{-t} \end{smallmatrix}\right)$ (equivalently, any ergodic component of μ' with respect to the action of the group $a'(t) = (a(t), e)$) has positive entropy (in fact, $h_{\mu_x^{\mathcal{E}}}(a) \geq 2/9$). One can now use Theorem 2.9 for $\mathrm{PGL}(2,\mathbb{R})$ to deduce that μ' and hence the arithmetic quantum limit μ is proportional to the Haar measure.

The entropy bound of [6] is proved by giving a uniform bound for every x in a compact $K \subset \Gamma\backslash\mathrm{PGL}(2,\mathbb{R})$ on the measure of geometric tubes around x of the form $\mu(B_k.x) < c_{B,K}\exp(-4k/9)$, with B_k as in §3.2.4. This bound holds already for appropriate lifts of the measures $\tilde{\mu}_n$, and depends only on ϕ_n being eigenfunctions of all Hecke operators (but does not use that they are also eigenfunctions of the Laplacian).

5.2.7. Using the same general strategy, L. Silberman and A. Venkatesh have been able to prove arithmetic quantum unique ergodicity for other $\Gamma\backslash G/K$, specifically for $G = \mathrm{PGL}(p, \mathbb{R})$ ($p > 2$ prime) and Γ a lattice arising from an order in a division algebra of degree p over $\mathbb{Q}$ (for $p = 2$ this gives precisely the lattices considered in part 2 of §5.2.1). While the strategy remains the same, several new ideas are needed for this extension, in particular a new micro-local lift for higher rank groups [68], [67]. Silberman has informed us that these methods can be used to establish analogs of Theorem 5.2 in some three-dimensional hyperbolic (arithmetic) manifolds; to the best of our knowledge such extensions are not known for any higher dimensional hyperbolic manifolds.

5.3. A result of N. Anantharaman. We conclude this section by discussing some very recent results of N. Anantharaman which shed some light on quantum limits in full generality (and not just in the arithmetic context). It can be deduced from her paper [1] that if M is a compact manifold with negative sectional curvature then every quantum limit has positive ergodic theoretic entropy. In the case of surfaces of constant curvature -1 Anantharaman actually proves that for any $\delta > 0$ any quantum limit has a positive measure of ergodic components with ergodic theoretic entropy $\geq (d-1)/2 - \delta$; in this normalization the ergodic theoretic entropy of the uniform measure m_{SM} is $d - 1$.

An exposition of some of her ideas as well as a different but closely related approach, both applied to a simpler toy model, can be found in [2] by N. Anantharaman and S. Nonnenmacher.

Note that in contrast to [6], this method inherently can only prove that some ergodic components have positive entropy. In the nonarithmetic situation it seems very hard to show that all ergodic components have positive entropy; indeed, this is false in the toy model considered in [2] as well as for an appropriately quantized hyperbolic toral automorphism [23].

6. Entropy and distribution of periodic orbits

Let G be a semisimple $\mathbb{R}$-split algebraic group, $\Gamma < G$ a lattice, and A an $\mathbb{R}$-split Cartan subgroup of G [53]. Then there are always infinitely many periodic A orbits in $\Gamma\backslash G$. Uniform measure on these orbits give examples of A-invariant measures with zero entropy. It is surprising therefore that entropy plays a key role in our understanding of distribution properties of such compact orbits. The results described in this section are joint work of the authors with P. Michel and A. Venkatesh [21] and are also described from a somewhat different viewpoint in [50] in these proceedings. Unless otherwise stated, proofs of all the statements below can be found in [21]. Periodic orbits of $\mathbb{R}$-split Cartan subgroups have also been studied elsewhere. We mention in particular the papers [52] by H. Oh where finiteness theorems regarding these orbits are proved and [4] where Y. Benoist and H. Oh prove equidistribution of Hecke orbits of a fixed A-periodic orbit.

6.1. Discriminant and regulators of periodic orbits. For concreteness, we restrict to the case $G = \mathrm{PGL}(n, \mathbb{R})$, $\Gamma = \mathrm{PGL}(n, \mathbb{Z})$ and $A < G$ the group of diagonal matrices. Later, we will also allow Γ to be a lattice associated to an order in a division algebra D over $\mathbb{Q}$ of degree n with $D \otimes \mathbb{R} \cong M(n, \mathbb{R})$ (e.g. for $n = 2$ a lattice as in part 2 in §5.2.1).

6.1.1. We wish to attach to every periodic A orbit in $X_n = \Gamma\backslash G$ two invariants: discriminant and regulator. Before doing this, we recall the following classical construction of such orbits:

Let K be a totally real extension of $\mathbb{Q}$ with $[K : \mathbb{Q}] = n$. Let $\mathcal{O}_K$ be the integers of K, and let $I \lhd \mathcal{O}_K$ be an ideal. Chose an ordering $\tau_1, \dots, \tau_n$ of the n embeddings of K in $\mathbb{R}$, and let $\tau = (\tau_1, \dots, \tau_n) \colon K \to \mathbb{R}^n$. Then $\tau(I)$ is a lattice in $\mathbb{R}^n$, hence corresponds to a point $(\!(g_I)\!) \in X_n$. If $\alpha_1, \dots, \alpha_n$ generate I as an additive group we can take $g_I = (\tau_j(\alpha_i))_{i,j=1}^n$. For any $v \in \mathbb{R}^n$ we let $\mathrm{diag}(v)$ be the diagonal matrix with entries $v_1, v_2, \dots, v_n$. If $\alpha \in \mathcal{O}_K^*$ then $I = \alpha I$ and

$$(\!(g_I)\!) = \mathrm{diag}(\tau(\alpha)).(\!(g_I)\!).$$

$\mathrm{diag}(\tau(\cdot))$ embeds $\mathcal{O}_K^*$ discretely in A, and Dirichlet's unit theorem gives us that $\mathcal{O}_K^*/\{\pm 1\}$ is a free Abelian group with $n-1$ generators. It follows that $(\!(g_I)\!)$ has periodic orbit under A. Two ideals $I, J \lhd \mathcal{O}_K$ are equivalent if $I = \alpha J$ for some $\alpha \in K$; in this case $\mathrm{diag}(\tau(\alpha)).(\!(g_I)\!) = (\!(g_J)\!)$ hence $A.(\!(g_I)\!) = A.(\!(g_J)\!)$ iff $I \sim J$. The number of equivalence classes of ideals is denoted by C_K and is called the *class number* of K; given τ we see that there are C_K distinct periodic A-orbits associated with $\mathcal{O}_K$.

A small variation of this construction allowing a general order $\mathcal{O} < \mathcal{O}_K$ and I a proper ideal of $\mathcal{O}$ gives all periodic A-orbits.

6.1.2. Let D be the algebra of all $n \times n$ (not necessarily invertible) diagonal matrices over $\mathbb{R}$. Let $x = (\!(g)\!)$ be a point with periodic A-orbit. Define $\Gamma_{A,g} := \Gamma \cap gAg^{-1}$. Then since x is periodic, $g^{-1}\Gamma_{A,g}g$ is a lattice in A. To $\Gamma_{A,g}$ we can also attach a subring $\mathbb{Q}[\Gamma_{A,g}]$ in $M(n, \mathbb{R})$ in an obvious way. Let $\Delta_{A,g} = g^{-1}(\mathbb{Q}[\Gamma_{A,g}] \cap M(n, \mathbb{Z}))g$; this ring can be shown to be a lattice in D (considered as an additive group).

We define the *regulator* $\mathrm{reg}(x)$ of a periodic $x \in X_n$ for $x = (\!(g)\!)$ to be the volume of $A/(g^{-1}\Gamma_{A,g}g)$ and the *discriminant* $\mathrm{disc}(x)$ by

$$\mathrm{disc}(x) = m_D(D/\Delta_{A,g})^2.$$

Note that $\mathrm{reg}(x)$ is simply the volume of the periodic orbit $A.x$. This defines both of these notions up to a global multiplicative constant corresponding to fixing a Haar measure on A and D respectively. This normalization can be chosen in such a way that $\mathrm{disc}(x)$ is an integer for every A-periodic x and so that for $(\!(g_I)\!)$, g_I as in §6.1.1 for $I \lhd \mathcal{O}_K$, the discriminant and regulator of $(\!(g_I)\!)$ coincide with the discriminant

and regulator of the number field K. The regulator and discriminant of a periodic orbit are related, but in a rather weak way: in general for any periodic orbit $A.x$,

$$\log \operatorname{disc}(x) \ll \operatorname{reg}(x) \ll_\varepsilon \operatorname{disc}(x)^{1/2+\varepsilon}.$$

If e.g. K has no subfields other than $\mathbb{Q}$ the lower bound on $\operatorname{reg}(x)$ can be improved to $c'_n[\log \operatorname{disc}(x)]^{n-1}$, which is tight. However, if K, I and $x = (\!(g_I)\!)$ are as in §6.1.1, $C_K \operatorname{reg}(x)$, i.e. the total volume of all periodic A orbits coming from ideals in $\mathcal{O}_K$, is closely related to the discriminant:

$$\operatorname{disc}(x)^{1/2-o(1)} \leq C_K \operatorname{reg}(x) \leq \operatorname{disc}(x)^{1/2+o(1)}.$$

Properly formulated (the easiest formulation is Adelic) a similar relation holds also for periodic A orbits coming from non-maximal orders $\mathcal{O} < \mathcal{O}_K$.

6.2. Some distribution properties of periodic orbits. We would like to prove statements regarding how sequences of periodic orbits are distributed. Care must be taken however as even for the simplest cases such as X_3 it is not true that for any sequence of A-periodic points $x_i \in X_3$ with $\operatorname{disc}(x_i) \to \infty$ the orbits $A.x_i$ become equidistributed. Let $\mu_{A.x_i}$ be the unique A-invariant probability measure on $A.x_i$. One problem, for $n = 3$ or more generally, is that it is possible to construct sequences x_i such that $\mu_{A.x_i}$ converge weak* to a measure μ with $\mu(X_n) < 1$. However, as remarked by Margulis the following is a consequence of Conjecture 4.4:

Conjecture 6.1. For any fixed compact $K \subset X_n$, $n \geq 3$, there are only finitely many periodic A-orbits contained in K.

Using Corollary 2.3 and what we call the Linnik principle (see §6.3) we prove the following towards this conjecture:

Theorem 6.2 ([21]). *For any fixed compact $K \subset X_n$, $n \geq 3$, for any $\varepsilon > 0$, the total volume of all periodic A-orbits contained in K of discriminant $\leq D$ is at most $O_\varepsilon(D^\varepsilon)$.*

In contrast, for $n = 2$, for any ε one can find a compact $K_\varepsilon \subset X_2$ so that the total volume of all periodic A-orbits contained in K of discriminant $\leq D$ is $\gg D^{1-\varepsilon}$.

Theorem 6.2 directly implies that for any $\varepsilon > 0$ and $n \geq 3$ the number of totally real numbers fields K of degree $[K : \mathbb{Q}] = n$ and discriminant $\operatorname{disc}(K) \leq D$ for which for some ideal class there is no representative of norm $\leq \varepsilon \operatorname{disc}(K)^{1/2}$ is $\ll D^\varepsilon$, giving a partial answer to the third question posed in the introduction.

The same method gives the following in the compact case:

Theorem 6.3 ([21]). *Let Γ be a lattice in $\mathrm{PGL}(n, \mathbb{R})$ associated with a division algebra over degree n over $\mathbb{Q}$ and $\eta > 0$ arbitrary. For any i let $(x_{i,j})_{j=1,\dots,N_i}$ be a finite collection of A-periodic points with distinct A-orbits such that*

$$\sum_{j=1}^{N_i} \operatorname{reg}(x_{i,j}) \geq \max_j (\operatorname{disc}(x_{i,j}))^\eta.$$

Suppose that there is no locally homogeneous proper subset of $\Gamma\backslash G$ *containing infinitely many* $x_{i,j}$. *Then* $\overline{\bigcup_{i,j} A.x_{i,j}} = \Gamma\backslash G$.

6.3. The Linnik principle. In the proofs of Theorems 6.2 and 6.3 a crucial point is establishing that a limiting measure has positive entropy under some $a \in A$, which allows one to apply the measure classification results described in §2. Positivity of the entropy is established using the following proposition, which links entropy with the size (regulator) of a periodic orbit (or a collection of periodic orbits) compared to its discriminant(s). We give it below for some lattices Γ in $G = \mathrm{PGL}(n, \mathbb{R})$, but this phenomenon is much more general. We call this relation between orbit size and entropy the *Linnik principle* in honor of Yu. Linnik in whose book [43] a special case of this relation is implicit.

Proposition 6.4 ([21]). *For every* ℓ, *let* $A.x_{\ell,j}$ $(j = 1 \dots N_\ell)$ *be a finite collection of* (*distinct*) *periodic* A*-orbits in* $\Gamma\backslash \mathrm{PGL}(n, \mathbb{R})$ *with* Γ *either* $\mathrm{PGL}(n, \mathbb{Z})$ *or a lattice corresponding to an order in a division algebra of degree* n *over* $\mathbb{Q}$. *Let* $\mu_{(\ell)}$ *be the average of the measures* $\mu_{A.x_{\ell,1}}, \dots, \mu_{A.x_{\ell,N_\ell}}$ *weighted by regulator. Suppose that*

$$\sum_{j=1}^{N_\ell} \mathrm{reg}(x_{\ell,j}) \geq \max_j (\mathrm{disc}(x_{\ell,j}))^\eta.$$

and that the $\mu_{(\ell)}$ *converge weak* to a* probability *measure* μ. *Then for any regular*[19] $a \in A$, *there is an* $c_{a,n} > 0$ (*which can be easily made explicit*) *so that*

$$h_\mu(a) \geq c_{a,n}\eta. \tag{6.1}$$

If $\mathrm{disc}(x_{\ell,j}) = \mathrm{disc}(x_{\ell,j'})$ *for all* ℓ, j, j' *then* (6.1) *can be improved to* $h_\mu(a) \geq 2c_{a,n}\eta$.

The key observation lies in the fact that for any compact $K \subset \Gamma\backslash G$ if x and $x' \in K$ are periodic with discriminant $\leq D$ then either $x = a.x'$ for some small $a \in A$ or $d(x, x') > CD^{-2}$ where $C = C(K)$ depends on K (if $\mathrm{disc}(x) = \mathrm{disc}(x')$ then $d(x, x') > CD^{-1}$ hence the improved estimate in this case). This fact is used together with the subadditivity of $H_{\mu_{(\ell)}}(\cdot)$ (see §3.1.1) to show that $h_\mu(a) \geq c_{a,n}\eta$.

6.4. Packets of periodic orbits and Duke's theorem. Another completely different way to establish positive entropy for limiting measures is given by subconvex estimates on L-functions.

Theorem 6.5 ([21]). *Let* $K_{(\ell)}$ *be a sequence of totally real degree three extensions of* $\mathbb{Q}$, $C_{(\ell)}$ *the class number of* $K_{(\ell)}$, *and* $\tau_{(\ell)} : K_{(\ell)} \to \mathbb{R}^3$ *a* 3*-tuple of embeddings. Let* $A.x_{1,\ell}, \dots, A.x_{C_{(\ell)},\ell}$ *be the periodic* A *orbits corresponding to the ideal classes of* $K_{(\ell)}$ *as in* §6.1.1. *Let* $\mu_{(\ell)} = \frac{1}{C_{(\ell)}} \sum_i \mu_{A.x_{i,\ell}}$. *Then* $\mu_{(\ell)}$ *converge in the weak* topology to the* $\mathrm{PGL}(3, \mathbb{R})$ *invariant probability measure* m_{X_3} *on* X_3.

[19] I.e. with no multiple eigenvalues.

Subconvexity estimates of W. Duke, J. Friedlander and H. Iwaniec [11] imply that for certain test functions f, the integrals $\int_{X_3} f d\mu_{(\ell)}$ converge to the right value (i.e. $\int_{X_3} f dm_{X_3}$). The space of test functions on which this convergence can be established from the subconvex estimates of Duke, Friedlander and Iwaniec is far from dense, but is sufficiently rich to show that any limiting measure of $\mu_{(\ell)}$ is a probability measure (i.e. there is no escape of mass to the cusp) and that the entropy of *every* ergodic component in such a limiting measure is greater than an explicit lower bound. Once these two facts have been established, Theorem 2.2 can be used to bootstrap entropy to equidistribution.

Theorem 6.5 is a generalization to the case $n = 3$ of the following theorem of Duke, proved using earlier and related subconvexity estimates of Duke and Iwaniec:

Theorem 6.6 (Duke [10]). *Let $K_{(\ell)} = \mathbb{Q}(\sqrt{D_\ell})$ be a sequence of real quadratic fields and $\mu_{(\ell)}$ an average of the corresponding measures on A-periodic orbits in X_2 as above. Then $\mu_{(\ell)}$ converge weak* to m_{X_2}.*

We note that Duke also gives an explicit rate of equidistribution of the $\mu_{(\ell)}$.

There is an alternative, ergodic theoretic, approach to this theorem that dates back to Yu. Linnik and B. F. Skubenko. Skubenko [70], building on work of Linnik [43], used this approach, which is closely related to techniques discussed in §6.3, to prove Theorem 6.6 under a congruence condition on the sequence D_ℓ (see [43, Chapter VI]). In [22] we show that a variation of this method can actually be used to give a complete proof of Theorem 6.6 using only ergodic theory and some properties of quadratic forms.

References

[1] Anantharaman, Nalini, Entropy and the localization of eigenfunctions. Preprint, 2004.

[2] Anantharaman, Nalini, and Nonnenmacher, Stéphane, Entropy of semiclassical measures of the Walsh-quantized baker's map. Preprint, 2005.

[3] Barnett, Alex, Asymptotic rate of quantum ergodicity in chaotic euclidean billiards. Preprint, 2004.

[4] Benoist, Yves, and Oh, Hee, Equidistribution of rational matrices in their conjugacy classes. *Geom. Funct. Anal.*, to appear.

[5] Berend, Daniel, Multi-invariant sets on compact abelian groups. *Trans. Amer. Math. Soc.* **286** (2) (1984), 505–535.

[6] Bourgain, Jean, and Lindenstrauss, Elon, Entropy of quantum limits. *Comm. Math. Phys.* **233** (1) (2003), 153–171.

[7] Brin, M., and Katok, A., On local entropy. In *Geometric dynamics* (Rio de Janeiro, 1981), Lecture Notes in Math. 1007, Springer-Verlag, Berlin 1983, 30–38.

[8] Cassels, J. W. S., and Swinnerton-Dyer, H. P. F., On the product of three homogeneous linear forms and the indefinite ternary quadratic forms. *Philos. Trans. Roy. Soc. London. Ser. A.* **248** (1955), 73–96.

[9] Colin de Verdière, Y., Ergodicité et fonctions propres du laplacien. *Comm. Math. Phys.* **102** (3) (1985), 497–502.

[10] Duke, W., Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.* **92** (1) (1988), 73–90.

[11] Duke, W., Friedlander, J. B., and H. Iwaniec, H., The subconvexity problem for Artin L-functions. *Invent. Math.* **149** (3) (2002), 489–577.

[12] Einsiedler, Manfred, and Katok, Anatole, Invariant measures on G/Γ for split simple Lie groups G. *Comm. Pure Appl. Math.* **56** (8) (2003), 1184–1221. Dedicated to the memory of Jürgen K. Moser.

[13] Einsiedler, Manfred, and Katok, Anatole, Rigidity of measures – the high entropy case, and non-commuting foliations. *Israel J. Math.* **148** (2005), 169–238.

[14] Einsiedler, Manfred, Katok, Anatole, and Lindenstrauss, Elon, Invariant measures and the set of exceptions to Littlewood's conjecture. *Ann. of Math.*, to appear.

[15] Einsiedler, Manfred, and Kleinbock, Dmitry, Measure Rigidity and p-adic Littlewood type problems. Preprint, 2005.

[16] Einsiedler, Manfred, and Lindenstrauss, Elon, Rigidity properties of Z^d-actions on tori and solenoids. *Electron. Res. Announc. Amer. Math. Soc.* **9** (2003), 99–110.

[17] Einsiedler, Manfred, and Lindenstrauss, Elon, Joining of higher rank diagonalizable actions on locally homogeneous spaces. Submitted, 2006.

[18] Einsiedler, Manfred, and Lindenstrauss, Elon, Joining of higher rank diagonalizable actions on locally homogeneous spaces (II). In preparation, 2006.

[19] Einsiedler, Manfred, and Lindenstrauss, Elon, On measures invariant under maximal split tori for semisimple S-algebraic groups. In preparation, 2006.

[20] Einsiedler, Manfred, and Lindenstrauss, Elon, Rigidity of measures invariant under a diagonalizable group – the general low entropy method. In preparation, 2006.

[21] Einsiedler, Manfred, Lindenstrauss, Elon, Michel, Philippe, and Venkatesh, Akshay, Distribution properties of compact torus orbits on homogeneous spaces. In preparation, 2006.

[22] Einsiedler, Manfred, Lindenstrauss, Elon, Michel, Philippe, and Venkatesh, Akshay, Distribution of compact torus orbits II. In preparation, 2006.

[23] Faure, Frédéric, Nonnenmacher, Stéphane, and De Bièvre, Stephan, Scarred eigenstates for quantum cat maps of minimal periods. *Comm. Math. Phys.* **239** (3) (2003), 449–492.

[24] Feldman, J., A generalization of a result of R. Lyons about measures on [0, 1). *Israel J. Math.* **81** (3) (1993), 281–287.

[25] Glasner, Eli, *Ergodic theory via joinings*. Math. Surveys Monogr. 101, Amer. Math. Soc., Providence, RI, 2003.

[26] Host, Bernard, Nombres normaux, entropie, translations. *Israel J. Math.* **91** (1–3) (1995), 419–428.

[27] Host, Bernard, Some results of uniform distribution in the multidimensional torus. *Ergodic Theory Dynam. Systems* **20** (2) (2000), 439–452.

[28] Jakobson, Dmitri, Equidistribution of cusp forms on $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathrm{PSL}_2(\mathbb{R})$. *Ann. Inst. Fourier* (*Grenoble*) **47** (3) (1997), 967–984.

[29] Johnson, Aimee S. A., Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers. *Israel J. Math.* **77** (1–2) (1992), 211–240.

[30] Kalinin, Boris, and Katok, Anatole, Invariant measures for actions of higher rank abelian groups. In *Smooth ergodic theory and its applications* (Seattle, WA, 1999), Proc. Sympos. Pure Math. 69, Amer. Math. Soc., Providence, RI, 2001, 593–637.

[31] Kalinin, Boris, and Katok, Anatole, Measurable rigidity and disjointness for $\mathbb{Z}^k$ actions by toral automorphisms. *Ergodic Theory Dynam. Systems* **22** (2) (2002), 507–523.

[32] Kalinin, Boris, and Spatzier, Ralf, Rigidity of the measurable structure for algebraic actions of higher-rank Abelian groups. *Ergodic Theory Dynam. Systems* **25** (1) (2005), 175–200.

[33] Katok, A., and Spatzier, R. J., Invariant measures for higher-rank hyperbolic abelian actions. *Ergodic Theory Dynam. Systems* **16** (4) (1996), 751–778.

[34] Katok, A., and Spatzier, R. J., Corrections to: "Invariant measures for higher-rank hyperbolic abelian actions" [Ergodic Theory Dynam. Systems **16** (4) (1996), 751–778]. *Ergodic Theory Dynam. Systems* **18** (2) (1998), 503–507.

[35] Katok, Anatole, and Hasselblatt, Boris, *Introduction to the modern theory of dynamical systems*. With a supplement by Anatole Katok and Leonardo Mendoza, Encyclopedia Math. Appl. 54, Cambridge University Press, Cambridge 1995.

[36] Katok, Anatole, Katok, Svetlana, and Schmidt, Klaus, Rigidity of measurable structure for $\mathbb{Z}^d$-actions by automorphisms of a torus. *Comment. Math. Helv.* **77** (4) (2002), 718–745.

[37] Kleinbock, Dmitry, Shah, Nimish, and Starkov, Alexander, Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In *Handbook of dynamical systems*, Vol. 1A, North-Holland, Amsterdam 2002, 813–930.

[38] Lindenstrauss, Elon, On quantum unique ergodicity for $\Gamma\backslash\mathbb{H} \times \mathbb{H}$. *Internat. Math. Res. Notices* **2001** (17) (2001), 913–933.

[39] Lindenstrauss, Elon, Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math.*, to appear.

[40] Lindenstrauss, Elon, Arithmetic quantum unique ergodicity and adelic dynamics. In *Proceedings of Current Developments in Mathematics conference* (Harvard, 2004), to appear.

[41] Lindenstrauss, Elon, and Venkatesh, Akshay, Existence and Weyl's law for spherical cusp forms. *Geom. Funct. Anal.*, to appear.

[42] Lindenstrauss, Elon, and Weiss, Barak, On sets invariant under the action of the diagonal group. *Ergodic Theory Dynam. Systems* **21** (5) (2001), 1481–1500.

[43] Linnik, Yu. V., *Ergodic properties of algebraic fields*. Translated from the Russian by M. S. Keane, Ergeb. Math. Grenzgeb. 45, Springer-Verlag, New York 1968.

[44] Luo, Wen Zhi, and Sarnak, Peter, Quantum ergodicity of eigenfunctions on $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}^2$. *Inst. Hautes Études Sci. Publ. Math.* **(81)** (1995), 207–237.

[45] Luo, Wen Zhi, and Sarnak, Peter, Quantum variance for Hecke eigenforms. Preprint, 2004.

[46] Lyons, Russell, On measures simultaneously 2- and 3-invariant. *Israel J. Math.* **61** (2) (1988), 219–224.

[47] Margulis, G. A., Oppenheim conjecture. In *Fields Medallists' lectures*, World Sci. Ser. 20th Century Math. 5, World Sci. Publishing, River Edge, NJ, 1997, 272–327.

[48] Margulis, G. A., and Tomanov, G. M., Invariant measures for actions of unipotent groups over local fields on homogeneous spaces. *Invent. Math.* **116** (1–3) (1994), 347–392.

[49] Margulis, Gregory, Problems and conjectures in rigidity theory. In *Mathematics: frontiers and perspectives*, Amer. Math. Soc., Providence, RI, 2000, 161–174.

[50] Michel, Philippe, and Venkatesh, Akshay, Equidistribution, L-functions and ergodic theory: on some problems of Y. V. Linnik. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume II, EMS Publishing House, Zürich 2006, 421–457.

[51] Morris, Dave Witte, *Ratner's Theorems on Unipotent Flows*. Chicago Lectures in Math., University of Chicago Press, Chicago, IL, 2005.

[52] Oh, Hee, Finiteness of compact maximal flats of bounded volume. *Ergodic Theory Dynam. Systems* **24** (1) (2004), 217–225.

[53] Prasad, Gopal, and Raghunathan, M. S., Cartan subgroups and lattices in semi-simple groups. *Ann. of Math. (2)* 96 (1972), 296–317.

[54] Ratner, Marina, Factors of horocycle flows. *Ergodic Theory Dynam. Systems* **2** (3–4) (1982), 465–489.

[55] Ratner, Marina, Rigidity of horocycle flows. *Ann. of Math.* (2) **115** (3) (1982), 597–614.

[56] Ratner, Marina, Horocycle flows, joinings and rigidity of products. *Ann. of Math.* (2) **118** (2) (1983), 277–313.

[57] Ratner, Marina, On Raghunathan's measure conjecture. *Ann. of Math.* (2) **134** (3) (1991), 545–607.

[58] Ratner, Marina,, Raghunathan's conjectures for Cartesian products of real and p-adic Lie groups. *Duke Math. J.* **77** (2) (1995), 275–382.

[59] Ratner, Marina,, Interactions between ergodic theory, Lie groups, and number theory. In *Proceedings of the International Congress of Mathematicians* (Zürich, 1994), Vol. 1, Birkhäuser, Basel 1995, 157–182.

[60] Rees, M., Some R^2-anosov flows. Unpublished, 1982.

[61] Rudnick, Zeév, and Sarnak, Peter, The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.* **161** (1) (1990), 195–213, 1994.

[62] Rudolph, Daniel J., $\times 2$ and $\times 3$ invariant measures and entropy. *Ergodic Theory Dynam. Systems* **10** (2) (1990), 395–406.

[63] Sarnak, Peter, Arithmetic quantum chaos. In *The Schur lectures* (Tel Aviv, 1992), Israel Math. Conf. Proc. 8, Bar-Ilan University, Ramat Gan 1995, 183–236.

[64] Sarnak, Peter, Estimates for Rankin-Selberg L-functions and quantum unique ergodicity. *J. Funct. Anal.* **184** (2) (2001), 419–453.

[65] Sarnak, Peter, Spectra of hyperbolic surfaces. *Bull. Amer. Math. Soc.* (*N.S.*) **40** (4) (2003), 441–478 (electronic).

[66] Selberg, A., Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc.* (*N.S.*) **20** (1956), 47–87.

[67] Silberman, Lior, Arithmetic quantum chaos on locally symmetric spaces. Ph.D. thesis, Princeton University, 2005.

[68] Silberman, Lior, and Venkatesh, Akshay, On quantum unique ergodicity for locally symmetric spaces I: a micro local lift. Preprint, 2004.

[69] Šnirel'man. Ergodic properties of eigenfunctions. *Uspehi Mat. Nauk* **29** (6 (180)) (1974), 181–182.

[70] B. F. Skubenko, A. I., The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems. *Izv. Akad. Nauk SSSR Ser. Mat.* **26** (1962), 721–752.

[71] Tomanov, George, Values of decomposable forms at S-integer points and tori orbits on homogeneous spaces. Preprint, 2005.

[72] Tomanov, George, and Weiss, Barak, Closed orbits for actions of maximal tori on homogeneous spaces. *Duke Math. J.* **119** (2) (2003), 367–392.

[73] Watson, Thomas, Rankin triple products and quantum chaos. Ph.D. thesis, Princeton University, 2001.

[74] Wolpert, Scott A., The modulus of continuity for $\Gamma_0(m)\backslash\mathbb{H}$ semi-classical limits. *Comm. Math. Phys.* **216** (2) (2001), 313–323.

[75] Zelditch, Steven, Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* **55** (4) (1987), 919–941.

Department of Mathematics, Ohio State University, Columbus, OH 43210, U.S.A.

Department of Mathematics, Princeton University, Princeton, NJ 08544, U.S.A.

Author index